

A SIMPLE REWARD-FREE APPROACH TO CONSTRAINED REINFORCEMENT LEARNING

Anonymous authors

Paper under double-blind review

ABSTRACT

In constrained reinforcement learning (RL), a learning agent seeks to not only optimize the overall reward but also satisfy the additional safety, diversity, or budget constraints. Consequently, existing constrained RL solutions require several new algorithmic ingredients that are notably different from standard RL. On the other hand, reward-free RL is independently developed in the unconstrained literature, which learns the transition dynamics without using the reward information, and thus naturally capable of addressing RL with multiple objectives under the common dynamics. This paper bridges reward-free RL and constrained RL. Particularly, we propose a simple meta-algorithm such that given any reward-free RL oracle, the approachability and constrained RL problems can be directly solved with negligible overheads in sample complexity. Utilizing the existing reward-free RL solvers, our framework provides sharp sample complexity results for constrained RL in the tabular MDP setting, matching the best existing results up to a factor of horizon dependence; our framework directly extends to a setting of tabular two-player Markov games, and gives a new result for constrained RL with linear function approximation.

1 INTRODUCTION

In a wide range of modern reinforcement learning (RL) applications, it is not sufficient for the learning agents to only maximize a scalar reward. More importantly, they must satisfy various *constraints*. For instance, such constraints can be the physical limit of power consumption or torque in motors for robotics tasks (Tessler et al., 2019); the budget for computation and the frequency of actions for real-time strategy games (Vinyals et al., 2019); and the requirement for safety, fuel efficiency and human comfort for autonomous drive (Le et al., 2019). In addition, constraints are also crucial in tasks such as dynamic pricing with limited supply (Besbes & Zeevi, 2009; Babaioff et al., 2015), scheduling of resources on a computer cluster (Mao et al., 2016), imitation learning (Syed & Schapire, 2007; Ziebart et al., 2008; Sun et al., 2019), as well as RL with fairness (Jabbari et al., 2017).

These huge demand in practice gives rise to a subfield—constrained RL, which focuses on designing efficient algorithms to find near-optimal policies for RL problems under linear or general convex constraints. Most constrained RL works directly combine the existing techniques such as value iteration and optimism from unconstrained literature, with new techniques specifically designed to deal with linear constraints (Efroni et al., 2020; Ding et al., 2021; Qiu et al., 2020) or general convex constraints (Brantley et al., 2020; Yu et al., 2021). The end product is a single new complex algorithm which is tasked to solve all the challenges of learning dynamics, exploration, planning as well as constraints satisfaction simultaneously. Thus, these algorithms need to be re-analyzed from scratch, and it is highly nontrivial to translate the progress in the unconstrained RL to the constrained setting.

On the other hand, reward-free RL—proposed in Jin et al. (2020a)—is a framework for the unconstrained setting, which learns the transition dynamics without using the reward. The framework has two phases: in the exploration phase, the agent first collects trajectories from a Markov decision process (MDP) and learns the dynamics without a pre-specified reward function. After exploration, the agent is tasked with computing near-optimal policies under the MDP for a collection of given reward functions. This framework is particularly suitable when there are multiple reward functions of interest, and has been developed recently to attack various settings including tabular MDPs (Jin et al.,

Table 1: Sample complexity for algorithms to solve reward-free RL for VMDP (Definition 1), approachability (Definition 3) and CMDP with general convex constraints (Definition 4).¹

	Algorithm	Reward-free	Approachability	CMDP
Tabular	Wu et al. (2020)	$\tilde{O}(\min\{d, S\}H^4SA/\epsilon^2)$	-	-
	Brantley et al. (2020)	-	-	$\tilde{O}(d^2H^3S^2A/\epsilon^2)$
	Yu et al. (2021)	-	$\tilde{O}(\min\{d, S\}H^3SA/\epsilon^2)$	$\tilde{O}(\min\{d, S\}H^3SA/\epsilon^2)$
	This work	$\tilde{O}(\min\{d, S\}H^4SA/\epsilon^2)$	$\tilde{O}(\min\{d, S\}H^4SA/\epsilon^2)$	$\tilde{O}(\min\{d, S\}H^4SA/\epsilon^2)$
Linear	This work	$\tilde{O}(d_{\text{lin}}^3H^6/\epsilon^2)$	$\tilde{O}(d_{\text{lin}}^3H^6/\epsilon^2)$	$\tilde{O}(d_{\text{lin}}^3H^6/\epsilon^2)$

2020a; Zhang et al., 2020), linear MDPs (Wang et al., 2020; Zanette et al., 2020), and tabular Markov games (Liu et al., 2020).

Contribution. In this paper, we propose a simple approach to solve constrained RL problems by bridging the reward-free RL literature and constrained RL literature. Our approach isolates the challenges of constraint satisfaction, and leaves the remaining RL challenges such as learning dynamics and exploration to reward-free RL. This allows us to design a new algorithm which purely focuses on addressing the constraints. Formally, we design a meta-algorithm for RL problems with general convex constraints. Our meta-algorithm takes a reward-free RL solver, and can be used to directly solve the approachability problem, as well as the constrained MDP problems using very small amount of samples in addition to what is required for reward-free RL.

Our framework enables direct translation of any progress in reward-free RL to constrained RL. Leveraging recent advances in reward-free RL, our meta-algorithm directly implies sample-efficient guarantees of constrained RL in the settings of tabular MDP, linear MDP, as well as tabular two-player Markov games. In particular,

- *Tabular setting:* Our work achieves sample complexity of $\tilde{O}(\min\{d, S\}H^4SA/\epsilon^2)$ for all three tasks of reward-free RL for Vector-valued MDPs (VMDP), approachability, and RL with general convex constraints. Here d is the dimension of VMDP or the number of constraints, S, A are the number of states and actions, H is the horizon, and ϵ is the error tolerance. It matches the best existing results up to a factor of H .
- *Linear setting:* Our work provides new sample complexity of $\tilde{O}(d_{\text{lin}}^3H^6/\epsilon^2)$ for all three tasks above for linear MDPs. To our best knowledge, this result is the first sample-efficient result for approachability and also constrained RL with general convex constraints in the linear function approximation setting.
- *Two-player setting:* Our work extends to the setting of tabular two-player vector-valued Markov games and achieves low regret of $\alpha(T) = \mathcal{O}(\epsilon/2 + \sqrt{H^2\iota/T})$ at the cost of this $\mathcal{O}(\epsilon)$ bias in regret as well as additional samples for preprocessing.

1.1 RELATED WORK

In this section, we review the related works on three tasks studied in this paper—reward-free RL, approachability, and constrained RL.

Reward-free RL. Reward-free exploration has been formalized by Jin et al. (2020a) for the tabular setting. Furthermore, Jin et al. (2020a) proposed an algorithm which has sample complexity $\tilde{O}(\text{poly}(H)S^2A/\epsilon^2)$ outputting ϵ -optimal policy for arbitrary number of reward functions. More

¹The presented sample complexities are all under the L_2 normalization conditions as studied in this paper. We comment that the results of (Wu et al., 2020; Brantley et al., 2020; Yu et al., 2021) are originally presented under L_1/L_∞ normalization conditions. While the results in Wu et al. (2020) can be directly adapted to our setting as stated in the table, the other two results Brantley et al. (2020); Yu et al. (2021) will be no better than the displayed results after adaptation.

recently, Zhang et al. (2020); Liu et al. (2020) propose algorithm VI-Zero with sharp sample complexity of $\tilde{O}(\text{poly}(H) \log(N)SA/\epsilon^2)$ capable of handling N fixed reward functions. Wang et al. (2020); Zanette et al. (2020) further provide reward-free learning results in the setting of linear function approximation, in particular, Wang et al. (2020) guarantees to find the near-optimal policies for an arbitrary number of (linear) reward functions within a sample complexity of $\tilde{O}(\text{poly}(H)d_{\text{lin}}^3/\epsilon^2)$. All results mentioned above are for scalar-valued MDPs. For the vector-valued MDPs (VMDPs), very recent work of Wu et al. (2020) designs a reward-free algorithm with sample complexity guarantee $\tilde{O}(\text{poly}(H) \min\{d, S\}SA/\epsilon^2)$ in the tabular setting. Compared to Wu et al. (2020), our reward-free algorithms for VMDP is adapted from the VI-Zero algorithm presented in Liu et al. (2020); While achieving the same sample complexity, it allows arbitrary planning algorithms in the planning phase.

Approachability and Constrained RL Approachability and Constrained RL are two related tasks involving constraints. Inspired by Blackwell approachability (Blackwell et al., 1956), recent work of Miryosefi et al. (2019) introduces approachability task for VMDPs. However, the proposed algorithm does not have polynomial sample complexity guarantees. More recently, Yu et al. (2021) gave a new algorithm for approachability for both VMDPs and vector-valued Markov games (VMGs). Yu et al. (2021) provides regret bounds for the proposed algorithm resulting in sample complexity guarantees of $\tilde{O}(\text{poly}(H) \min\{d, S\}SA/\epsilon^2)$ for approachability in VMDPs and $\tilde{O}(\text{poly}(H) \min\{d, S\}SAB/\epsilon^2)$ for approachability in VMGs.

Sample-efficient exploration in constrained reinforcement learning has been recently studied in a recent line of work by Brantley et al. (2020); Qiu et al. (2020); Efroni et al. (2020); Ding et al. (2021); Singh et al. (2020). All these works are also limited to linear constraints except Brantley et al. (2020) which extends their approach to general convex constraints achieving sample complexity of $\tilde{O}(\text{poly}(H)d^2S^2A/\epsilon^2)$. However, Brantley et al. (2020) requires solving a large-scale convex optimization sub-problem. The best result for constrained RL with general convex constraints can be achieved by the approachability-based algorithm in Yu et al. (2021) obtaining sample complexity of $\tilde{O}(\text{poly}(H) \min\{d, S\}SA/\epsilon^2)$. Technically, our meta-algorithm is based on the Fenchel’s duality, which is similar to Yu et al. (2021). In contrast, Yu et al. (2021) does not use reward-free RL, and is thus different from our results in terms of algorithmic approaches. Consequently, Yu et al. (2021) does not reveal the deep connections between reward-free RL and constrained RL, which is one of the main contribution of this paper. In addition, Yu et al. (2021) does not address the function approximation setting.

Finally, we note that among all results mentioned above, only Ding et al. (2021) has considered models beyond tabular setting in the context of constrained RL. The model studied in Ding et al. (2021) is known as linear mixture MDPs which is different and incomparable to the linear MDP models considered in this paper. We further comment that Ding et al. (2021) can only handle linear constraints for CMDP, while our results is capable of solving CMDPs with general convex constraints.

2 PRELIMINARIES AND PROBLEM SETUP

We consider an episodic *vector-valued Markov decision process* (VMDP) specified by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \mathbf{r})$, where \mathcal{S} is the state space, \mathcal{A} is the action space, H is the length of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ is the collection of *unknown* transition probabilities with $\mathbb{P}_h(s' | s, a)$ equal to the probability of transiting to s' after taking action a in state s at the h^{th} step, and $\mathbf{r} = \{\mathbf{r}_h : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{B}(1)\}_{h=1}^H$ is a collection of *unknown* d -dimensional return functions, where $\mathcal{B}(r)$ is the d -dimensional Euclidean ball of radius r centered at the origin.

Interaction protocol. In each episode, agent starts at a *fixed* initial state s_1 . Then, at each step $h \in [H]$, the agent observes the current state s_h , takes action a_h , receives stochastic sample of the return vector $\mathbf{r}_h(s_h, a_h)$, and it causes the environment to transit to $s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h)$. We assume that stochastic samples of the return function are also in $\mathcal{B}(1)$, almost surely.

Policy and value function. A policy π of an agent is a collection of H functions $\{\pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})\}_{h=1}^H$ that map states to distribution over actions. The agent following policy π , picks action $a_h \sim \pi_h(s_h)$ at the h^{th} step. We denote $\mathbf{V}_h^\pi : \mathcal{S} \rightarrow \mathcal{B}(H)$ as the value func-

tion at step h for policy π , defined as $\mathbf{V}_h^\pi(s) := \mathbb{E}_\pi \left[\sum_{h'=h}^H \mathbf{r}_{h'}(s_{h'}, a_{h'}) \mid s_h = s \right]$. Similarly, we denote $\mathbf{Q}_h^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{B}(H)$ as the Q -value function at step h for policy π , where $\mathbf{Q}_h^\pi(s, a) := \mathbb{E}_\pi \left[\sum_{h'=h}^H \mathbf{r}_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a \right]$.

Scalarized MDP. For a VMDP \mathcal{M} and $\boldsymbol{\theta} \in \mathcal{B}(1)$, we define scalar-valued MDP $\mathcal{M}_\theta = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r_\theta)$, where $r_\theta = \{\langle \boldsymbol{\theta}, \mathbf{r}_h \rangle : \mathcal{S} \times \mathcal{A} \rightarrow [-1, 1]\}_{h=1}^H$. We denote $V_h^\pi(\cdot; \boldsymbol{\theta}) : \mathcal{S} \rightarrow [-H, H]$ as the scalarized value function at step h for policy π , defined as

$$V_h^\pi(s; \boldsymbol{\theta}) := \mathbb{E}_\pi \left[\sum_{h'=h}^H \langle \boldsymbol{\theta}, \mathbf{r}_{h'}(s_{h'}, a_{h'}) \rangle \mid s_h = s \right] = \langle \boldsymbol{\theta}, \mathbf{V}_h^\pi(s) \rangle.$$

Similarly, we denote $Q_h^\pi(\cdot; \boldsymbol{\theta}) : \mathcal{S} \times \mathcal{A} \rightarrow [-H, H]$ as the scalarized Q -value function at step h for policy π , where

$$Q_h^\pi(s, a; \boldsymbol{\theta}) := \mathbb{E}_\pi \left[\sum_{h'=h}^H \langle \boldsymbol{\theta}, \mathbf{r}_{h'}(s_{h'}, a_{h'}) \rangle \mid s_h = s, a_h = a \right] = \langle \boldsymbol{\theta}, \mathbf{Q}_h^\pi(s, a) \rangle.$$

For a fixed $\boldsymbol{\theta} \in \mathbb{R}^d$, there exists an optimal policy π_θ^* , maximizing value for all states (Puterman, 2014); i.e., $V_h^{\pi_\theta^*}(s; \boldsymbol{\theta}) = \sup_\pi V_h^\pi(s; \boldsymbol{\theta})$ for all $s \in \mathcal{S}$ and $h \in [H]$. We abbreviate $V^{\pi_\theta^*}(\cdot; \boldsymbol{\theta})$ and $Q^{\pi_\theta^*}(\cdot; \boldsymbol{\theta})$ as $V^*(\cdot; \boldsymbol{\theta})$ and $Q^*(\cdot; \boldsymbol{\theta})$ respectively.

2.1 REWARD-FREE EXPLORATION (RFE) FOR VMDPS

The task of *reward-free exploration* (formalized by Jin et al. (2020a) for tabular MDPs) considers the scenario in which the agents interacts with the environment without guidance of reward information. Later, the reward information is revealed and the agents is required to compute the near-optimal policy. In this section, we describe its counterpart for VMDPs¹. Formally, it consists of two phases:

Exploration phase. In the exploration phase, agent explores the unknown environment without observing any information regarding the return function. Namely, at each episode the agent executes policies to collect samples. The policies can depend on dynamic observations $\{s_h^k, a_h^k\}_{(k,h) \in [K] \times [H]}$ in the past episodes, but not the return vectors.

Planning phase. In the planning phase, the agent no longer interacts with the environment; however, stochastic samples of the d -dimensional return function for the collected episodes is revealed to the agent, i.e. $\{\mathbf{r}_h^k\}_{(k,h) \in [K] \times [H]}$. Based on the episodes collected during the exploration phase, the agent outputs the near-optimal policies of \mathcal{M}_θ given an arbitrary number of vectors $\boldsymbol{\theta} \in \mathcal{B}(1)$.

Definition 1 (Reward-free algorithm for VMDPs). For any $\epsilon, \delta > 0$, after collecting $m_{\text{RFE}}(\epsilon, \delta)$ episodes during the exploration phase, with probability at least $1 - \delta$, the algorithm satisfies

$$\forall \boldsymbol{\theta} \in \mathcal{B}(1) : V_1^*(s_1; \boldsymbol{\theta}) - V_1^{\pi_\theta}(s_1; \boldsymbol{\theta}) \leq \epsilon, \quad (1)$$

where π_θ is the output of the planning phase for vector $\boldsymbol{\theta}$ as input. The function m_{RFE} determines the *sample complexity* of the RFE algorithm.

Remark 2. *Standard reward-free setup concerns MDPs with scalar reward, and requires the algorithm to find the near-optimal policies for N different prespecified reward functions in the planning phase, where the sample complexity typically scales with $\log N$. This type of results can be adapted into a guarantee in the form of (1) for VMDP by ϵ -covering of $\boldsymbol{\theta}$ over $\mathcal{B}(1)$ and a modified concentration arguments (see the proofs of Theorem 7 and Theorem 13 for more details).*

2.2 APPROACHABILITY FOR VMDPS

In this section we provide the description for the *approachability* task for VMDPs introduced by Miryoosefi et al. (2019). Given a vector-valued Markov decision process and a convex target set \mathcal{C} , the goal is to learn a policy whose expected cumulative return vector lies in the target set (akin to Blackwell approachability in single-turn games, Blackwell et al. 1956). We consider the agnostic version of this task which is more general since it doesn't need to assume that such policy exists; instead, the agent learns to minimize the Euclidean distance between expected return of the learned policy and the target set.

¹RFE for VMDPs is also called preference-free exploration problem in Wu et al. (2020)

Definition 3 (Approachability algorithm for VMDPs). For any $\epsilon, \delta > 0$, after collecting $m_{\text{APP}}(\epsilon, \delta)$ episodes, with probability at least $1 - \delta$, the algorithm satisfies

$$\text{dist}(\mathbf{V}_1^{\pi^{\text{out}}}(s_1), \mathcal{C}) \leq \min_{\pi} \text{dist}(\mathbf{V}_1^{\pi}(s_1), \mathcal{C}) + \epsilon, \quad (2)$$

where π^{out} is the output of the algorithm and $\text{dist}(\mathbf{x}, \mathcal{C})$ is the Euclidean distance between point \mathbf{x} and set \mathcal{C} . The function m_{APP} determines the *sample complexity* of the algorithm.

2.3 CONSTRAINED MDP (CMDP) WITH GENERAL CONVEX CONSTRAINTS

In this section we describe *constrained Markov decision processes* (CMDPs) introduced by [Altman \(1999\)](#). The goal of this setting is to minimize cost while satisfying some linear constraints over consumption of d resources (resources are akin to \mathbf{r} in our case). Although, the original definition only allows for linear constraints, we consider the more general case of arbitrary convex constraints. More formally, consider a VMDP \mathcal{M} , a cost function $c = \{c_h : \mathcal{S} \times \mathcal{A} \rightarrow [-1, 1]\}_{h=1}^H$, and a convex constraint set \mathcal{C} . The agent goal is to compete against the following benchmark:

$$\min_{\pi} C_1^{\pi}(s_1) \quad \text{s.t.} \quad \mathbf{V}_1^{\pi}(s_1) \in \mathcal{C},$$

where $C_h^{\pi} = \mathbb{E}_{\pi} \left[\sum_{h'=h}^H c_{h'}(s_{h'}, a_{h'}) \mid s_h = s \right]$.

Definition 4 (Algorithm for CMDP). For any $\epsilon, \delta > 0$, after collecting $m_{\text{CMDP}}(\epsilon, \delta)$ episodes, with probability at least $1 - \delta$, the algorithm satisfies

$$\begin{cases} C_1^{\pi^{\text{out}}}(s_1) - \min_{\pi: \mathbf{V}_1^{\pi}(s_1) \in \mathcal{C}} C_1^{\pi}(s_1) \leq \epsilon \\ \text{dist}(\mathbf{V}_1^{\pi^{\text{out}}}(s_1), \mathcal{C}) \leq \epsilon, \end{cases} \quad (3)$$

where π^{out} is the output of the algorithm. The function m_{CMDP} determines the *sample complexity* of the algorithm.

As also mentioned in the prior works ([Miryoosefi et al., 2019](#); [Yu et al., 2021](#)), we formally show in the following theorem that approachability task (Definition 3) can be considered more general compared to CMDP (Definition 4); Namely, given any algorithm for the former we can obtain an algorithm for the latter by incurring only extra logarithmic factor and a negligible overhead. The idea is to incorporate cost into the constraint set \mathcal{C} and perform an (approximate) binary search over the minimum attainable cost. The reduction and the proof can be found in [Appendix A](#).

Theorem 5. *Given any approachability algorithm (Definition 3) with sample complexity m_{APP} , we can design an algorithm for CMDP (Definition 4) with sample complexity m_{CMDP} , satisfying*

$$m_{\text{CMDP}}(\epsilon, \delta) \leq \mathcal{O} \left(m_{\text{APP}} \left(\frac{\epsilon}{6}, \frac{\epsilon \delta}{12H} \right) + \frac{H^2 \log[dH/\epsilon\delta]}{\epsilon^2} \right) \cdot \log \frac{1}{\epsilon}.$$

3 META-ALGORITHM FOR VMDPS

In this section, equipped with preliminaries discussed in [Section 2](#), we are ready to introduce our main algorithmic framework for VMDPs bridging reward-free RL and approachability.

Before introducing the algorithm, we explain the intuition behind it. By Fenchel’s duality (similar to [Yu et al. 2021](#)), one can show that

$$\min_{\pi} \text{dist}(\mathbf{V}_1^{\pi}(s_1), \mathcal{C}) = \min_{\pi} \max_{\theta \in \mathcal{B}(1)} \left[\langle \theta, \mathbf{V}_1^{\pi}(s_1, \mathcal{C}) \rangle - \max_{\mathbf{x}' \in \mathcal{C}} \langle \theta, \mathbf{x}' \rangle \right].$$

It satisfies the minimax conditions since it’s concave in θ and convex in π (by allowing mixture policies); therefore, minimax theorem [Neumann \(1928\)](#) implies that we can equivalently solve

$$\max_{\theta \in \mathcal{B}(1)} \min_{\pi} \left[\langle \theta, \mathbf{V}_1^{\pi}(s_1, \mathcal{C}) \rangle - \max_{\mathbf{x}' \in \mathcal{C}} \langle \theta, \mathbf{x}' \rangle \right].$$

This max-min form allows us to use general technique of [Freund & Schapire \(1999\)](#) for solving a max-min by repeatedly playing a no-regret online learning algorithm as the max-player against best-response for the min-player. In particular, for a fixed θ , minimizing over π is equivalent to finding optimal policy for scalarized MDP $\mathcal{M}_{-\theta}$. To achieve this, we can utilize a reward-free oracle as in [Definition 1](#). On the other hand for θ -player we are able to use online gradient descent ([Zinkevich, 2003](#)). By combining ideas above, we obtain [Algorithm 1](#).

Algorithm 1 Meta-algorithm for VMDPs

```

1: Input: Reward-Free Algorithm RFE for VMDPs (as in Definition 1), Target Set  $\mathcal{C}$ 
2: Hyperparameters: learning rate  $\eta^t$ 
3: Initialize: run exploration phase of RFE for  $K$  episodes
4: Set:  $\theta^1 \in \mathcal{B}(1)$ 
5: for  $t = 1, 2, \dots, T$  do
6:   Obtain near optimal policy for  $\mathcal{M}_{-\theta^t}$ :
        $\pi^t \leftarrow$  output of planning phase of RFE for preference vector  $-\theta^t$ 
7:   Estimate  $\mathbf{V}_1^{\pi^t}(s_1)$  using one episode:
       Run  $\pi^t$  for one episode and let  $\hat{\mathbf{v}}^t$  be the sum of vectorial returns
8:   Apply online gradient ascent update for utility function  $u^t(\theta) = \langle \theta, \hat{\mathbf{v}}^t \rangle - \max_{\mathbf{x} \in \mathcal{C}} \langle \theta, \mathbf{x} \rangle$ :
        $\theta^{t+1} \leftarrow \Gamma_{\mathcal{B}(1)}[\theta^t + \eta^t(\hat{\mathbf{v}}^t - \arg\max_{\mathbf{x} \in \mathcal{C}} \langle \theta^t, \mathbf{x} \rangle)]$ 
       where  $\Gamma_{\mathcal{B}(1)}$  is the projection into Euclidean unit ball
9: Let  $\pi^{\text{out}}$  be uniform mixture of  $\{\pi^1, \dots, \pi^T\}$ 
10: Return  $\pi^{\text{out}}$ 

```

Theorem 6. *There exists an absolute constant c , such that for any choice of RFE algorithm (Definition 1) and for any $\epsilon \in (0, H]$ and $\delta \in (0, 1]$, if we choose*

$$T \geq c(H^2\iota/\epsilon^2), \quad K \geq m_{\text{RFE}}(\epsilon/2, \delta/2), \quad \text{and} \quad \eta^t = \sqrt{1/(H^2t)},$$

where $\iota = \log(d/\delta)$; then, with probability at least $1 - \delta$, Algorithm 1 outputs an ϵ -optimal policy for the approachability (Equation 2). Therefore, we have $m_{\text{APP}}(\epsilon, \delta) \leq \mathcal{O}(m_{\text{RFE}}(\epsilon/2, \delta/2) + H^2\iota/\epsilon^2)$.

Theorem 6 shows that given any reward-free algorithm, Algorithm 1 can solve the approachability task with negligible overhead. The proof for Theorem 6 is provided in Appendix B. Equipped with this theorem, since we have already shown the connection between approachability and constrained RL in Theorem 5, any results for RFE can be directly translated to results for constrained RL.

4 TABULAR VMDPS

In this section, we consider tabular VMDPs; namely, we assume that $|\mathcal{S}| \leq S$ and $|\mathcal{A}| \leq A$. Utilizing prior work on tabular setting, we describe our choice of reward-free algorithm.

In the exploration phase, we use VI-Zero proposed by Liu et al. (2020). It can be seen as UCB-VI (Azar et al., 2017) with zero reward. Intuitively, the value function computed in the algorithm measures the level of uncertainty and incentivizes the greedy policy to visit underexplored states. The output of VI-Zero is $\hat{\mathbb{P}}^{\text{out}}$, which is an estimation of the transition dynamics.

In the planning phase, given $\theta \in \mathcal{B}(1)$ we can use any planning algorithm (e.g., value iteration) for $\widehat{\mathcal{M}}_\theta = (\mathcal{S}, \mathcal{A}, H, \hat{\mathbb{P}}^{\text{out}}, \langle \theta, \hat{\mathbf{r}} \rangle)$ where $\hat{\mathbf{r}}$ is empirical estimate of \mathbf{r} using collected samples $\{\mathbf{r}_h^k\}$.

The following theorem state theoretical guarantees for tabular VMDPs. Proof of Theorem 7 and more details can be found in Appendix C.

Theorem 7. *For tabular VMDP, we have a reward-free algorithm (Definition 1) with $m_{\text{RFE}}(\epsilon, \delta) \leq \mathcal{O}(\min\{d, S\}H^4SA\iota/\epsilon^2 + H^3S^2A\iota^2/\epsilon)$, an algorithm for approachability (Definition 3) with $m_{\text{APP}}(\epsilon, \delta) \leq \mathcal{O}(\min\{d, S\}H^4SA\iota/\epsilon^2 + H^3S^2A\iota^2/\epsilon)$, and an algorithm for CMDP (Definition 4) with $m_{\text{CMDP}}(\epsilon, \delta) \leq \mathcal{O}(\min\{d, S\}H^4SA\iota^2/\epsilon^2 + H^3S^2A\iota^3/\epsilon)$.*

The reward-free algorithm with stated sample complexity in Theorem 7 is the VI-Zero algorithm (Algorithm 5 in Appendix C). Its sample complexity result is obtained by adapting the results in Liu et al. (2020) for scalar-valued MDPs to the settings of VMDPs in this paper. The algorithms for approachability and CMDP is based on plugging in VI-Zero into our meta algorithms, and the corresponding sample complexity results are obtained by applying Theorem 5 and our main result—Theorem 6.

Theorem 7 shows that the sample complexity of all three tasks are connected—the leading terms are all $\tilde{O}(\min\{d, S\}H^4SA/\epsilon^2)$ which differ by only logarithmic factors. In particular, our sample complexity for the reward-free exploration (Definition 1) in the tabular setting matches the best result in Wu et al. (2020). It further shows that we can easily design an sample-efficient for approachability (Definition 3) and CMDP with general convex constraints (Definition 4) in the tabular setting, with sample complexity matching the best result in Yu et al. (2021) up to a single factor of H .² Therefore, our framework while being modular enabling direct translation of reward-free RL to constrained RL, achieves sharp sample complexity guarantees. We comment that due to reward-free nature of our approach unlike Yu et al. (2021), we can no longer provide regret guarantees.

5 LINEAR FUNCTION APPROXIMATION: LINEAR VMDPS

In this section we consider the setting of linear function approximation and allow \mathcal{S} and \mathcal{A} to be infinitely large. We assume that agent has access to a feature map $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^{d_{\text{lin}}}$ and the return function and transitions are linear functions of the feature map. We formally define the linear VMDPs in Assumption 8 which adapts the definition of linear MDPs (Jin et al., 2020b) for VMDPs; namely, they coincide for the case of $d = 1$.

Assumption 8 (Linear VMDP). A VMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \mathbf{r})$ is said to be a linear VMDP with a feature map $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^{d_{\text{lin}}}$, if for any $h \in [H]$:

1. There exists d_{lin} unknown (signed) measures $\boldsymbol{\mu}_h = \{\mu_h^{(1)}, \dots, \mu_h^{(d_{\text{lin}})}\}$ over \mathcal{S} such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ we have $\mathbb{P}_h(\cdot | s, a) = \langle \boldsymbol{\mu}(\cdot), \phi(s, a) \rangle$.
2. There exists an unknown matrix $W_h \in \mathbb{R}^{d \times d_{\text{lin}}}$ such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ we have $\mathbf{r}_h(s, a) = W_h \phi(s, a)$.

Similar to Jin et al. (2020b), we assume that $\|\phi(s, a)\| \leq 1$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\|\boldsymbol{\mu}_h(\mathcal{S})\| \leq \sqrt{d_{\text{lin}}}$ for all $h \in [H]$, and $\|W_h\| \leq \sqrt{d_{\text{lin}}}$ for all $h \in [H]$.

Wang et al. (2020) has recently proposed a sample-efficient algorithm for reward-free exploration in linear MDPs. Utilizing that algorithm and tailoring it for our setting, we can obtain the following theoretical guarantee. The algorithm and the proof can be found in Appendix D.

Theorem 9. For linear VMDPs (Assumption 8), we have a reward-free algorithm (Definition 1) with $m_{\text{RFE}}(\epsilon, \delta) \leq \mathcal{O}(d_{\text{lin}}^3 H^6 \iota^2 / \epsilon^2)$, an approachability algorithm (Definition 3) with $m_{\text{APP}}(\epsilon, \delta) \leq \mathcal{O}(d_{\text{lin}}^3 H^6 \iota^2 / \epsilon^2)$ and an algorithm for CMDP (Definition 4) with $m_{\text{CMDP}}(\epsilon, \delta) \leq \mathcal{O}(d_{\text{lin}}^3 H^6 \iota^3 / \epsilon^2)$.

The reward-free algorithm with stated sample complexity in Theorem 9 is the Algorithm 6 in Appendix D. It is a modified version of the reward-free algorithm introduced by Wang et al. (2020). Its sample complexity result is again obtained by adapting the results in Wang et al. (2020) for scalar-valued MDPs to the settings of VMDPs in this paper. The algorithms for approachability and CMDP is based on plugging in this reward-free algorithm into our meta algorithms, and the corresponding sample complexity results are obtained by applying Theorem 5 and our main result—Theorem 6.

Theorem 9 provides a new sample complexity result of $\tilde{O}(d_{\text{lin}}^3 H^6 / \epsilon^2)$ for the reward-free exploration (Definition 1) in the linear setting (Assumption 8). It further provides a new sample complexity result of $\tilde{O}(d_{\text{lin}}^3 H^6 / \epsilon^2)$ for both approachability (Definition 3) and CMDP (Definition 4) in the linear setting (Assumption 8). To best our knowledge, this is the first sample-efficient result for constrained RL problems with linear function approximation and general convex constraints.

6 VECTOR-VALUED MARKOV GAMES

6.1 MODEL AND PRELIMINARIES

Similar to Section 2, we consider an episodic *vector-valued Markov game* (VMG) specified by a tuple $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \mathbb{P}, \mathbf{r})$, where \mathcal{A} and \mathcal{B} are the action spaces for the min-player and max-player,

²This H factor difference is due the Bernstein-type bonus used in Yu et al. (2021), which can not be adapted to the reward-free setting.

respectively. The d -dimensional return function \mathbf{r} and the transition probabilities \mathbb{P} , now depend on the current state and the action of both players.

Interaction protocol. In each episode, we start at a *fixed* initial state s_1 . Then, at each step $h \in [H]$, both players observe the current state s_h , take their own actions $a_h \in \mathcal{A}$ and $b_h \in \mathcal{B}$ simultaneously, observe stochastic sample of the return vector $\mathbf{r}_h(s_h, a_h, b_h)$ along with their opponent’s action, and it causes the environment to transit to $s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)$. We assume that stochastic samples of the return function are also in $\mathcal{B}(1)$, almost surely.

Policy and value function. A policy μ of the min-player is a collection of H functions $\{\mu_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})\}_{h=1}^H$. Similarly, a policy ν of the max-player is a collection of H functions $\{\nu_h : \mathcal{S} \rightarrow \Delta(\mathcal{B})\}_{h=1}^H$. If the players are following μ and ν , we have $a_h \sim \mu(\cdot | s)$ and $b_h \sim \nu(\cdot | s)$ at the h^{th} step. We use $\mathbf{V}_h^{\mu, \nu} : \mathcal{S} \rightarrow \mathcal{B}(H)$ and $\mathbf{Q}_h^{\mu, \nu} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \rightarrow \mathcal{B}(H)$ to denote the value function and Q-value function at step h under policies μ and ν .

Scalarized markov game and Nash equilibrium. For a VMG \mathcal{G} and $\theta \in \mathcal{B}(1)$, we define scalar-valued Markov game $\mathcal{G}_\theta = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r_\theta)$, where $r_\theta = \{\langle \theta, \mathbf{r}_h \rangle : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \rightarrow [-1, 1]\}_{h=1}^H$. We use $V_h^{\mu, \nu}(\cdot; \theta)$ and $Q_h^{\mu, \nu}(\cdot, \cdot, \cdot; \theta)$ to denote value function and Q-value function of \mathcal{G}_θ , respectively. Note that we have $V_h^{\mu, \nu}(s; \theta) = \langle \theta, \mathbf{V}_h^{\mu, \nu}(s) \rangle$ and $Q_h^{\mu, \nu}(s, a, b; \theta) = \langle \theta, \mathbf{Q}_h^{\mu, \nu}(s, a, b) \rangle$.

For any policy of the min-player μ , there exists a *best-response* policy $\nu_\dagger(\mu)$ of the max-player; i.e. $V_h^{\mu, \nu_\dagger(\mu)}(s; \theta) = \max_\nu V_h^{\mu, \nu}(s; \theta)$ for all $(s, h) \in \mathcal{S} \times [H]$. We use $V^{\mu, \dagger}$ to denote $V^{\mu, \nu_\dagger(\mu)}$. Similarly, we can define $\mu_\dagger(\nu)$ and $V^{\dagger, \nu}$. We further know (Filar & Vrieze, 2012) that there exist policies (μ^*, ν^*) , known as *Nash equilibrium*, satisfying the following equation for all $(s, h) \in \mathcal{S} \times [H]$:

$$\min_\mu \max_\nu V_h^{\mu, \nu}(s; \theta) = V_h^{\mu^*, \dagger}(s; \theta) = V_h^{\mu^*, \nu^*}(s; \theta) = V_h^{\dagger, \nu^*}(s; \theta) = \max_\nu \min_\mu V_h^{\mu, \nu}(s; \theta)$$

In words, it means that no player can gain anything by changing her own policy. We abbreviate $V_h^{\mu^*, \nu^*}$ and $Q_h^{\mu^*, \nu^*}$ as V_h^* and Q_h^* .

6.1.1 REWARD-FREE EXPLORATION (RFE) FOR VMGS

Similar to Section 2.1, we can define RFE algorithm for VMGs. Similarly, it consists of two phases. In the exploration phase the it explores the environment without guidance of return function. Later, in the planning phase, given any $\theta \in \mathcal{B}(1)$, it requires to output near optimal Nash equilibrium for \mathcal{G}_θ .

Definition 10 (RFE algorithm for VMGs). For any $\epsilon, \delta > 0$, after collecting $m_{\text{RFE}}(\epsilon, \delta)$ episodes during the exploration phase, with probability at least $1 - \delta$, the algorithm for all $\theta \in \mathcal{B}(1)$, satisfies

$$V_1^{\mu_\theta, \dagger}(s_1; \theta) - V_1^{\dagger, \nu_\theta}(s_1; \theta) = [V_1^{\mu_\theta, \dagger}(s_1; \theta) - V_1^*(s_1; \theta)] + [V_1^*(s_1; \theta) - V_1^{\dagger, \nu_\theta}(s_1; \theta)] \leq \epsilon$$

where (μ_θ, ν_θ) is the output of the planning phase for vector θ as input. The function m_{RFE} determines the *sample complexity* of the RFE algorithm.

6.1.2 BLACKWELL APPROACHABILITY FOR VMGS

We assume we are given a VMG \mathcal{G} and a target set \mathcal{C} . The goal of the min-player is for the return vector to lie in the set \mathcal{C} while max-player wants the opposite. For the two-player vector-valued games it can be easily shown that the minimax theorem does no longer hold (see Section 2.1 of Abernethy et al. 2011). Namely, if for every policy of the max-player we have a response such that the return is in the set, we cannot hope to find a single policy for the min-player so that for every policy of the max-player the return vector lie in the set. However, approaching the set on average is possible.

Definition 11 (Blackwell approachability). We say the min-player is approaching the target \mathcal{C} with rate $\alpha(T)$, if for arbitrary sequence of max-player policies ν^1, \dots, ν^T , we have

$$\text{dist}(\frac{1}{T} \sum_{t=1}^T \mathbf{V}_1^{\mu^t, \nu^t}(s_1), \mathcal{C}) \leq \max_\nu \min_\mu \text{dist}(\mathbf{V}_1^{\mu, \nu}(s_1), \mathcal{C}) + \alpha(T).$$

6.2 META-ALGORITHM FOR VMGS

Similar to Section 3, we introduce our main algorithmic framework for VMGs bridging reward-free algorithm and Blackwell approachability in VMGs. The pseudo-code is displayed in Algorithm 2 and the theoretical guarantees are provided in Theorem 12. The proof can be found in Appendix E.

Algorithm 2 Meta-algorithm for VMGs

- 1: **Input:** Reward-Free Algorithm RFE for VMG (as in Definition 10), Target Set \mathcal{C}
 - 2: **Hyperparameters:** learning rate η^t
 - 3: **Initialize:** run exploration phase of RFE for K episodes
 - 4: **Set:** $\theta^1 \in \mathcal{B}(1)$
 - 5: **for** $t = 1, 2, \dots, T$ **do**
 - 6: Obtain near optimal Nash equilibrium for \mathcal{G}_{θ^t} :
 $(\mu^t, \omega^t) \leftarrow$ output of planning phase of RFE for the vector θ^t as input
 - 7: Play μ^t for one episode:
 Play μ^t against max-player playing arbitrary policy ν^t for one episode
 and let $\hat{\mathbf{v}}^t$ be the sum of vectorial returns
 - 8: Apply online gradient ascent update for utility function $u^t(\theta) = \langle \theta, \hat{\mathbf{v}}^t \rangle - \max_{\mathbf{x} \in \mathcal{C}} \langle \theta, \mathbf{x} \rangle$:
 $\theta^{t+1} \leftarrow \Gamma_{\mathcal{B}(1)}[\theta^t + \eta^t(\hat{\mathbf{v}}^t - \arg\max_{\mathbf{x} \in \mathcal{C}} \langle \theta^t, \mathbf{x} \rangle)]$
 where $\Gamma_{\mathcal{B}(1)}$ is the projection into Euclidean unit ball
-

Theorem 12. *For any choice of RFE algorithm (Definition 10) and for any $\epsilon \in (0, H]$ and $\delta \in (0, 1]$, if we choose $K = m_{\text{RFE}}(\epsilon/2, \delta/2)$ and $\eta^t = \sqrt{1/H^2 t}$; then, with probability at least $1 - \delta$, the min-player in Algorithm 2, satisfies Definition 11 with rate $\alpha(T) = \mathcal{O}(\epsilon/2 + \sqrt{H^2 \iota/T})$ where $\iota = \log(d/\delta)$. Therefore to obtain ϵ -optimality, the total sample complexity scales with $\mathcal{O}(m_{\text{RFE}}(\epsilon/2, \delta/2) + H^2 \iota/\epsilon^2)$.*

6.3 TABULAR VMGS

In this section, we consider tabular VMDPs; namely, we assume that $|\mathcal{S}| \leq S$, $|\mathcal{A}| \leq A$, and $|\mathcal{B}| \leq B$. Similar to Section 4, by utilizing VI-Zero (Liu et al., 2020) we can have the following theoretical guarantees. The algorithm and the proof can be found in Appendix E.

Theorem 13. *There exists a reward-free algorithm for tabular VMGs and a right choice of hyperparameters that satisfies Definition 10 with sample complexity $m_{\text{RFE}}(\epsilon, \delta) \leq \mathcal{O}(\min\{d, S\}H^4 SAB\iota/\epsilon^2 + H^3 S^2 AB\iota^2/\epsilon)$, where $\iota = \log[dSABH/(\epsilon\delta)]$.*

The theorem provides a new sample complexity result of $\tilde{\mathcal{O}}(\min\{d, S\}H^4 SAB\iota/\epsilon^2)$ for reward-free exploration in VMGs (Definition 10). It immediately follows from Theorem 13 and Theorem 12 that we can achieve total sample complexity of $\tilde{\mathcal{O}}(\min\{d, S\}H^4 SAB\iota/\epsilon^2)$ for Blackwell approachability in VMGs (Definition 11). Our rate for $\alpha(T)$ scales with $\tilde{\mathcal{O}}(\sqrt{\text{poly}(H)/T})$ while the results in Yu et al. (2021) has the rate of $\alpha(T)$ scaling with $\tilde{\mathcal{O}}(\sqrt{\text{poly}(H) \min\{d, S\}SA/T})$. However, we require initial phase of self-play for $K = \mathcal{O}(m_{\text{RFE}})$ episodes which is not needed by Yu et al. (2021).

7 CONCLUSION

This paper provides a meta algorithm that takes a reward-free RL solver, and convert it to an algorithm for solving constrained RL problems. Our framework enables the direct translation of any progress in reward-free RL to constrained RL setting. Utilizing existing reward-free solvers, our framework provides sharp sample complexity results for constrained RL in tabular setting (matching best existing results up to factor of horizon dependence), new results for the linear function approximation setting. Our framework further extends to tabular two-player vector-valued Markov games for solving Blackwell approachability problem.

REFERENCES

- Jacob Abernethy, Peter L Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 27–46. JMLR Workshop and Conference Proceedings, 2011.
- Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.
- Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.
- Moshe Babaioff, Shaddin Dughmi, Robert D. Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply. *TEAC*, 3(1):4, 2015. Special issue for *13th ACM EC*, 2012.
- Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.
- David Blackwell et al. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.
- Kianté Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16315–16326. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/bc6d753857fe3dd4275dff707dedf329-Paper.pdf>.
- Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3304–3312. PMLR, 13–15 Apr 2021. URL <http://proceedings.mlr.press/v130/ding21d.html>.
- Yonathan Efroni, Shie Mannor, and Matteo Pirodda. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.
- Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.
- Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.
- Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *International Conference on Machine Learning*, pp. 1617–1626. PMLR, 2017.
- Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.
- Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020a.
- Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020b.
- Hoang Minh Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. *CoRR*, abs/1903.08738, 2019. URL <http://arxiv.org/abs/1903.08738>.
- Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020.

- Hongzi Mao, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, pp. 50–56, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450346610. doi: 10.1145/3005745.3005750. URL <https://doi.org/10.1145/3005745.3005750>.
- Sobhan Miryoosefi, Kianté Brantley, Hal Daume III, Miro Dudik, and Robert E Schapire. Reinforcement learning with convex constraints. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14093–14102. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/873be0705c80679f2c71fbf4d872df59-Paper.pdf>.
- J v Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.
- Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual optimization: Stochastically constrained markov decision processes with adversarial losses and unknown transitions. *arXiv preprint arXiv:2003.00660*, 2020.
- Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.
- Rahul Singh, Abhishek Gupta, and Ness B Shroff. Learning in markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.
- Wen Sun, Anirudh Vemula, Byron Boots, and J Andrew Bagnell. Provably efficient imitation learning from observation alone. *arXiv preprint arXiv:1905.10948*, 2019.
- Umar Syed and Robert E. Schapire. A game-theoretic approach to apprenticeship learning. In *Proceedings of the 20th International Conference on Neural Information Processing Systems, NIPS’07*, pp. 1449–1456, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.
- Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=SkfrvsA9FX>.
- Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.
- Ruosong Wang, Simon S Du, Lin F Yang, and Ruslan Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020.
- Jingfeng Wu, Vladimir Braverman, and Lin F Yang. Accommodating picky customers: Regret bound and exploration complexity for multi-objective reinforcement learning. *arXiv preprint arXiv:2011.13034*, 2020.
- Tiancheng Yu, Yi Tian, Jingzhao Zhang, and Suvrit Sra. Provably efficient algorithms for multi-objective competitive rl. *arXiv preprint arXiv:2102.03192*, 2021.
- Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. *arXiv preprint arXiv:2008.07737*, 2020.
- Xuezhou Zhang, Adish Singla, et al. Task-agnostic exploration in reinforcement learning. *arXiv preprint arXiv:2006.09497*, 2020.
- Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.
- Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003.

A PROOF FOR SECTION 2

In this section we provide proofs and missing details for Section 2.

A.1 PROOF OF THEOREM 5

Consider the following algorithm which is performing an approximate version of binary search on the optimal cost. We use \oplus to denote vector concatenation.

Algorithm 3 Solving Constrained RL Using Approachability

```

1: Input: approachability algorithm APP
2: Hyperparameters:  $\epsilon' > 0$ 
3: Initialize:  $L \leftarrow 0$  and  $R \leftarrow H$ 
4: Define the augmented VMDP model
   
$$\bar{\mathbf{r}}_h(s, a) = \mathbf{r}_h(s, a) \oplus c_h(s, a) \quad \forall h \in [H]$$

   
$$\bar{\mathcal{M}} = \{\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \bar{\mathbf{r}}\}$$

5: for iteration  $t = 1, 2, \dots, T$  do
6:   Set  $\text{mid} = (R + L)/2$ 
7:   Define the target set for approachability
   
$$\bar{\mathcal{C}}^t = \{\mathbf{x} \oplus y \mid \mathbf{x} \in \mathcal{C}, y \leq \text{mid}\}$$

8:    $\pi^t \leftarrow$  output of APP algorithm for the model  $\bar{\mathcal{M}}$  with target set  $\bar{\mathcal{C}}^t$  using  $K_{\text{APP}}$  episodes.
9:    $\bar{\mathbf{v}}^t \leftarrow$  estimate  $\bar{\mathbf{V}}_1^{\pi^t}(s_1)$  using  $K_{\text{est}}$  episodes, where  $\bar{\mathbf{V}}$  is the value function for  $\bar{\mathcal{M}}$ .
10:  if  $\text{dist}(\bar{\mathbf{v}}^t, \mathcal{C}) \leq \epsilon'$  then
11:     $R \leftarrow \text{mid}$ 
12:  else
13:     $L \leftarrow \text{mid}$ 
14: Return  $\pi^T$ 

```

Theorem 14. For any choice of approachability algorithm (as in Definition 3) and for any $\epsilon, \delta > 0$, if we choose

$$T = \mathcal{O}[\log(H/\epsilon)], \quad K_{\text{APP}} = m_{\text{APP}}(\epsilon, \epsilon\delta/(2H)), \quad K_{\text{est}} = \mathcal{O}\left[\frac{H^2 \log(dH/\epsilon\delta)}{\epsilon^2}\right], \quad \epsilon' = \mathcal{O}(\epsilon)$$

then, with probability at least $1 - \delta$, Algorithm 3 satisfies

$$\begin{cases} C_1^{\pi^T}(s_1) - \min_{\pi: \mathbf{V}_1^\pi(s_1) \in \mathcal{C}} C_1^\pi(s_1) \leq \mathcal{O}(\epsilon), \\ \text{dist}(\mathbf{V}_1^{\pi^T}(s_1), \mathcal{C}) \leq \mathcal{O}(\epsilon). \end{cases}$$

Proof of Theorem 14. By definition 3, Lemma 36, and union bound; with probability at least $1 - \delta$, we have for all $t \in [T]$

$$\begin{aligned} \|\bar{\mathbf{v}}^t - \bar{\mathbf{V}}_1^{\pi^t}(s_1)\| &\leq \epsilon, \\ \text{dist}(\bar{\mathbf{V}}_1^{\pi^t}(s_1), \mathcal{C}) &\leq \min_{\pi} \text{dist}(\bar{\mathbf{V}}_1^\pi(s_1), \mathcal{C}) + \epsilon. \end{aligned} \tag{4}$$

We use L^t , R^t , and mid^t to denote values of L , R , and mid during t^{th} iteration. By choice of T we have

$$R^T - L^T \leq \epsilon. \tag{5}$$

Define $c^* = \min_{\pi: \mathbf{V}_1^\pi(s_1) \in \mathcal{C}} C_1^\pi(s_1)$ and let $\pi^* = \arg\min_{\pi: \mathbf{V}_1^\pi(s_1) \in \mathcal{C}} C_1^\pi(s_1)$. Let's consider these cases

- Case $\text{mid} \geq c^*$: It's easy to see that $\min_{\pi} \text{dist}(\bar{\mathbf{V}}_1^\pi(s_1), \mathcal{C}) = 0$, therefore by second inequality in Equation 4 we have

$$\text{dist}(\bar{\mathbf{V}}_1^{\pi^t}(s_1), \mathcal{C}) \leq \epsilon.$$

Since distance function is 1-Lipschitz with respect to Euclidean norm, by first inequality in Equation 4, we have

$$\text{dist}(\bar{\mathbf{v}}^t, \mathcal{C}) \leq \epsilon + \epsilon = 2\epsilon$$

- Case $\text{mid} \leq c^* - 3\epsilon$: It's easy to see that $\min_{\pi} \text{dist}(\bar{\mathbf{V}}_1^{\pi}(s_1), \mathcal{C}) \geq 3\epsilon$, therefore by definition of minimum we have

$$\text{dist}(\bar{\mathbf{V}}_1^{\pi^t}(s_1), \mathcal{C}) \geq 3\epsilon.$$

Since distance function is 1-Lipschitz with respect to Euclidean norm, by first inequality in Equation 4, we have

$$\text{dist}(\bar{\mathbf{v}}^t, \mathcal{C}) \geq 3\epsilon - \epsilon = 2\epsilon.$$

What we showed above implies that if we set $\epsilon' = 2\epsilon$, in all iterations $t \in [T]$ we have

$$L^t \leq c^*, \quad R^t \geq c^* - 3\epsilon.$$

Combining with Equation 5, we get

$$c^* - 4\epsilon \leq L^T \leq \text{mid}^T \leq R^T \leq c^* + \epsilon$$

Therefore we have,

$$\begin{aligned} & \max\{C_1^{\pi^T}(s_1) - \text{mid}^T, \text{dist}(\mathbf{V}_1^{\pi^T}(s_1), \mathcal{C})\} \\ & \leq \text{dist}(\bar{\mathbf{V}}_1^{\pi^T}(s_1), \mathcal{C}) \\ & \leq \min_{\pi} \text{dist}(\bar{\mathbf{V}}_1^{\pi}(s_1), \mathcal{C}) + \epsilon \\ & \leq \text{dist}(\bar{\mathbf{V}}_1^{\pi^*}(s_1), \mathcal{C}) + \epsilon \\ & \leq \max\{c^* - \text{mid}^T, 0\} + \epsilon \\ & \leq c^* - (c^* - 4\epsilon) + \epsilon = 5\epsilon \end{aligned}$$

It implies

$$\begin{cases} \text{dist}(\mathbf{V}_1^{\pi^T}(s_1), \mathcal{C}) \leq 5\epsilon \\ C_1^{\pi^T}(s_1) \leq 5\epsilon + \text{mid}^T \leq c^* + 6\epsilon \end{cases}$$

Rescaling ϵ to $\epsilon/6$ completes the proof. \square

Proof of Theorem 5. Using Theorem 14 the claim follows immediately: total sample complexity of Algorithm 3 is

$$T(K_{\text{APP}} + K_{\text{est}}) \leq \log(1/\epsilon) \cdot \mathcal{O}\left(m_{\text{APP}}(\epsilon, \epsilon\delta/H) + \frac{H^2 \log[d/\epsilon\delta]}{\epsilon^2}\right).$$

\square

B PROOF FOR SECTION 3

In this section we provide proofs and missing details for Section 3.

B.1 FENCHEL DUALITY

Consider a convex and closed function $f : \text{dom}(f) \rightarrow \mathbb{R}$. We define the dual function f^* , called *Fenchel conjugate*, as

$$f^*(\theta) = \max_{\mathbf{x} \in \text{dom}(f)} [\langle \theta, \mathbf{x} \rangle - f(\mathbf{x})].$$

If function f is 1-Lipschitz and $\text{dom}(f) = \mathcal{B}(H)$; then, the conjugate function f^* is H -Lipschitz with $\text{dom}(f^*) = \mathcal{B}(1)$ (Corollary 13.3.3 in Rockafellar 2015). Therefore, Fenchel duality implies

$$f(\mathbf{x}) = \max_{\theta \in \mathcal{B}(1)} [\langle \theta, \mathbf{x} \rangle - f^*(\theta)].$$

In particular, for closed, convex, and 1-Lipschitz function f defined as

$$\begin{cases} f : \mathcal{B}(H) \rightarrow \mathbb{R} \\ f(\mathbf{x}) = \text{dist}(\mathbf{x}, \mathcal{C}) \end{cases}$$

we have

$$f^*(\boldsymbol{\theta}) = \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle.$$

It's easy to verify that $\partial f^*(\boldsymbol{\theta}) = \text{argmax}_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle$ is a subgradient of f^* at $\boldsymbol{\theta}$. Fenchel duality implies that

$$\text{dist}(x, \mathcal{C}) = \max_{\boldsymbol{\theta} \in \mathcal{B}(1)} \left[\langle \boldsymbol{\theta}, \mathbf{x} \rangle - \max_{\mathbf{x}' \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x}' \rangle \right]. \quad (6)$$

B.2 ONLINE CONVEX OPTIMIZATION (OCO)

We will be using the guarantee of online gradient ascent algorithm (Zinkevich, 2003) in the proof. Therefore, we briefly review the framework of online convex optimization. We can imagine an online game between the learner and the environment: The learner is given a decision set Θ ; at time $t = 1, 2, \dots, T$, the learner makes a decision $\boldsymbol{\theta}^t \in \Theta$, the environment reveals a concave utility function $u^t : \Theta \rightarrow \mathbb{R}$, and the learner gains utility $u^t(\boldsymbol{\theta}^t)$. The learner's goal is to minimize *regret* defined as

$$\text{Regret}_T \triangleq \max_{\boldsymbol{\theta} \in \Theta} \left[\sum_{t=1}^T u^t(\boldsymbol{\theta}) \right] - \left[\sum_{t=1}^T u^t(\boldsymbol{\theta}^t) \right].$$

An OCO algorithm is *no-regret* if $\text{Regret}_T = o(T)$, meaning its average utility approaches to best in hindsight. The *online gradient ascent* (OGA) is an example of such algorithm (Algorithm 4). In Theorem 15 we formally state the theoretical guarantee of this algorithm.

Algorithm 4 Online gradient ascent (OGA)

- 1: **input:** projection operator Γ_Θ where $\Gamma_\Theta(\boldsymbol{\theta}) = \text{argmin}_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$
 - 2: **init:** $\boldsymbol{\theta}^1$ arbitrarily
 - 3: **parameters:** step size η_t
 - 4: **for** $t = 1$ **to** T **do**
 - 5: observe concave utility function $u^t : \Theta \rightarrow \mathbb{R}$
 - 6: $\boldsymbol{\theta}^{t+1} = \Gamma_\Theta(\boldsymbol{\theta}^t + \eta_t \partial u^t(\boldsymbol{\theta}^t))$ {where $\partial u^t(\boldsymbol{\theta}^t)$ is a subgradient of u^t at $\boldsymbol{\theta}^t$ }
-

Theorem 15 (Zinkevich 2003). Assume that for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ we have $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \leq D$ and u^1, \dots, u^T are concave and G -Lipschitz. By setting $\eta_t = \frac{D}{G\sqrt{t}}$, Algorithm 4 satisfies

$$\text{Regret}_T \leq \mathcal{O}(DG\sqrt{T}).$$

B.3 PROOF OF THEOREM 6

We use the following choice for parameters:

$$K \geq m_{\text{RFE}}(\epsilon/2, \delta/2), \quad T \geq c \cdot (H^2 \iota / \epsilon^2). \quad (7)$$

We denote $\mathbf{v}^t := \mathbf{V}_1^{\pi^t}(s_1)$ and start with the following lemma.

Lemma 16. Define even E_0 to be:

$$\begin{cases} \left\| \frac{1}{T} \sum_{t=1}^T \mathbf{v}^t - \hat{\mathbf{v}}^t \right\| \leq \mathcal{O}(\sqrt{H^2 \iota / T}), \\ \mathbf{V}_1^*(s_1; -\boldsymbol{\theta}^t) \leq \mathbf{V}_1^{\pi^t}(s_1; -\boldsymbol{\theta}^t) + \epsilon/2 \quad \forall t \in [T]. \end{cases}$$

where $\iota = \log(d/\delta)$. We have $\mathbb{P}(E_0) \geq 1 - \delta$.

Proof of Lemma 16. We show that each claim holds with probability at least $1 - \delta/2$; applying a union bound completes the proof.

First claim. Let \mathcal{F}_t be the filtration capturing all the randomness in the algorithm before iteration t . We have $\mathbb{E}[\hat{\mathbf{v}}^t \mid \mathcal{F}_t] = \mathbf{v}^t$ and we also know that $\|\hat{\mathbf{v}}^t\| \leq H$ almost surely. By applying Lemma 36, with probability at least $1 - \delta$ we have

$$\left\| \frac{1}{T} \sum_{t=1}^T \mathbf{v}^t - \hat{\mathbf{v}}^t \right\| \leq \mathcal{O}(\sqrt{H^2 \log[d/\delta]/T}),$$

which completes the proof.

Second claim. Choice of parameters in Equation 7 along with Definition 1 immediately implies that with probability at least $1 - \delta/2$ we have

$$\mathbf{V}_1^*(s_1; -\boldsymbol{\theta}^t) \leq \mathbf{V}_1^{\pi^t}(s_1; -\boldsymbol{\theta}^t) + \epsilon/2 \quad \forall t \in [T].$$

Note that in Algorithm 1, π^t is the output of the planning phase of the RFE algorithm for the vector $-\boldsymbol{\theta}^t$ as input. \square

The following lemma states that if $\alpha = \min_{\pi} \text{dist}(\mathbf{V}_1^{\pi}(s_1), \mathcal{C}) \geq 0$ is the closest achievable distance to target set \mathcal{C} , then any halfspace containing \mathcal{C} is reachable up to error α .

Lemma 17. For any $\boldsymbol{\theta} \in \mathcal{B}(1)$, we have

$$\min_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle \leq \min_{\pi} \text{dist}(\mathbf{V}_1^{\pi}(s_1), \mathcal{C}) + \mathbf{V}_1^*(s_1; \boldsymbol{\theta}).$$

Proof of Lemma 17. Let $\bar{\pi} = \text{argmin}_{\pi} \text{dist}(\mathbf{V}_1^{\pi}(s_1), \mathcal{C})$ and define $\bar{\mathbf{v}} = \mathbf{V}_1^{\bar{\pi}}(s_1)$. Let $\tilde{\mathbf{v}} = \Gamma_{\mathcal{C}}(\bar{\mathbf{v}})$ be the orthogonal projection of $\bar{\mathbf{v}}$ into \mathcal{C} . We have

$$\begin{aligned} \mathbf{V}_1^*(s_1; \boldsymbol{\theta}) &\geq \mathbf{V}_1^{\bar{\pi}}(s_1; \boldsymbol{\theta}) \\ &= \langle \boldsymbol{\theta}, \bar{\mathbf{v}} \rangle \\ &= \langle \boldsymbol{\theta}, \bar{\mathbf{v}} - \tilde{\mathbf{v}} \rangle + \langle \boldsymbol{\theta}, \tilde{\mathbf{v}} \rangle \\ &\geq -\|\bar{\mathbf{v}} - \tilde{\mathbf{v}}\| + \min_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle \\ &\geq -\min_{\pi} \text{dist}(\mathbf{V}_1^{\pi}(s_1), \mathcal{C}) + \min_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle \end{aligned}$$

\square

Now we are ready to proceed with proof of Theorem 6.

Proof of Theorem 6. With probability at least $1 - \delta$ event E_0 holds and we have

$$\begin{aligned}
\text{dist}(\mathbf{V}_1^{\pi^{\text{out}}}(s_1), \mathcal{C}) &= \text{dist}\left(\frac{1}{T} \sum_{t=1}^T \mathbf{v}^t, \mathcal{C}\right) \\
&\stackrel{(i)}{=} \max_{\boldsymbol{\theta} \in \mathcal{B}(1)} \left[\langle \boldsymbol{\theta}, \frac{1}{T} \sum_{t=1}^T \mathbf{v}^t \rangle - \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle \right] \\
&= \max_{\boldsymbol{\theta} \in \mathcal{B}(1)} \left[\frac{1}{T} \sum_{t=1}^T (\langle \boldsymbol{\theta}, \hat{\mathbf{v}}^t \rangle - \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle) + \langle \boldsymbol{\theta}, \frac{1}{T} \sum_{t=1}^T \mathbf{v}^t - \hat{\mathbf{v}}^t \rangle \right] \\
&\stackrel{(ii)}{\leq} \max_{\boldsymbol{\theta} \in \mathcal{B}(1)} \left[\frac{1}{T} \sum_{t=1}^T (\langle \boldsymbol{\theta}, \hat{\mathbf{v}}^t \rangle - \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle) \right] + \mathcal{O}(\sqrt{H^2 \iota / T}) \\
&\stackrel{(iii)}{\leq} \frac{1}{T} \sum_{t=1}^T (\langle \boldsymbol{\theta}^t, \hat{\mathbf{v}}^t \rangle - \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}^t, \mathbf{x} \rangle) + \mathcal{O}(\sqrt{H^2 / T}) + \mathcal{O}(\sqrt{H^2 \iota / T}) \\
&\stackrel{(iv)}{\leq} \min_{\pi} \text{dist}(\mathbf{V}_1^{\pi}(s_1), \mathcal{C}) + \frac{1}{T} \sum_{t=1}^T (\langle \boldsymbol{\theta}^t, \hat{\mathbf{v}}^t \rangle + \mathbf{V}_1^*(s_1; -\boldsymbol{\theta}^t)) + \mathcal{O}(\sqrt{H^2 \iota / T}) \\
&\stackrel{(v)}{\leq} \min_{\pi} \text{dist}(\mathbf{V}_1^{\pi}(s_1), \mathcal{C}) + \epsilon/2 + \frac{1}{T} \sum_{t=1}^T (\langle \boldsymbol{\theta}^t, \hat{\mathbf{v}}^t \rangle + \mathbf{V}_1^{\pi^t}(s_1; -\boldsymbol{\theta}^t)) + \mathcal{O}(\sqrt{H^2 \iota / T}) \\
&= \min_{\pi} \text{dist}(\mathbf{V}_1^{\pi}(s_1), \mathcal{C}) + \epsilon/2 + \frac{1}{T} \sum_{t=1}^T \langle \boldsymbol{\theta}^t, \hat{\mathbf{v}}^t - \mathbf{v}^t \rangle + \mathcal{O}(\sqrt{H^2 \iota / T}) \\
&\stackrel{(vi)}{\leq} \min_{\pi} \text{dist}(\mathbf{V}_1^{\pi}(s_1), \mathcal{C}) + \epsilon/2 + \mathcal{O}(\sqrt{H^2 \iota / T}) \\
&\stackrel{(vii)}{\leq} \min_{\pi} \text{dist}(\mathbf{V}_1^{\pi}(s_1), \mathcal{C}) + \epsilon
\end{aligned}$$

where (i) is by Equation 6, (ii) is by first inequality in event E_0 together with Cauchy-Schwarz, (iii) is by guarantee of OGA in Theorem 15, (iv) is by Lemma 17, (v) is by second inequality in event E_0 , (vi) is by first inequality in event E_0 together with Cauchy-Schwarz, and finally (vii) is by setting $T \geq c(H^2 \iota / \epsilon^2)$ for large enough constant c , completing the proof. \square

C PROOF FOR SECTION 4

In this section we provide proofs and missing details for Section 4.

C.1 REWARD-FREE ALGORITHM FOR TABULAR VMDPs

In the exploration phase, we use VI-Zero (Liu et al., 2020) with modified choice of hyperparameters. The pseudocode is displayed in Algorithm 5. Intuitively, the value function $\tilde{Q}_h(s, a)$ computed in the algorithm measures the level of uncertainty that agent may suffer if it takes action a at state s in step h . It incentivize the greedy policy to visit underexplored states improving our empirical estimate $\hat{\mathbb{P}}$.

In the planning phase, given $\boldsymbol{\theta} \in \mathcal{B}(1)$ as input we can use any planning algorithm (such as value iteration) for $\hat{\mathcal{M}}_{\boldsymbol{\theta}} = (\mathcal{S}, \mathcal{A}, H, \hat{\mathbb{P}}^{\text{out}}, \langle \boldsymbol{\theta}, \hat{\mathbf{r}} \rangle)$ where $\hat{\mathbf{r}}$ is empirical estimate of \mathbf{r} using collected samples $\{\mathbf{r}_h^k\}$.

C.2 PROOF OF THEOREM 7

In this section, we prove Theorem 18 which implies the first claim in Theorem 7. Second and third claims in Theorem 7 immediately follow due to Theorem 6 and Theorem 5.

Let $\hat{\mathbb{P}}^k$ and $\hat{\mathbf{r}}^k$ be our empirical estimates of the transition and the return vectors at the beginning of the k^{th} episode in Algorithm 5 and define $\hat{\mathcal{M}}^k = (\mathcal{S}, \mathcal{A}, H, \hat{\mathbb{P}}^k, \hat{\mathbf{r}}^k)$. We use $N_h^k(s, a)$ to denote the

Algorithm 5 VI-Zero: Exploration Phase

```

1: Hyperparameters: Bonus  $\beta_t$ .
2: Initialize: for all  $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ :  $\tilde{Q}_h(s, a) \leftarrow H$  and  $N_h(s, a) \leftarrow 0$ ,
3:   for all  $(s, a, h, s') \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{S}$ :  $N_h(s, a, s') \leftarrow 0$ ,
4:    $\Delta \leftarrow 0$ .
5: for episode  $k = 1, 2, \dots, K$  do
6:   for step  $h = H, H-1, \dots, 1$  do
7:     for state-action pair  $(s, a) \in \mathcal{S} \times \mathcal{A}$  do
8:        $t \leftarrow N_h(s, a)$ .
9:       if  $t > 0$  then
10:         $\tilde{Q}_h(s, a) \leftarrow \min\{\hat{\mathbb{P}}_h \tilde{V}_{h+1}\}(s, a) + \beta_t, H\}$ .
11:       for state  $s \in \mathcal{S}$  do
12:         $\tilde{V}_h(s) \leftarrow \max_{a \in \mathcal{A}} \tilde{Q}_h(s, a)$  and  $\pi_h(s) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \tilde{Q}_h(s, a)$ 
13:       if  $\tilde{V}(s_1) \leq \Delta$  then
14:         $\Delta \leftarrow \tilde{V}(s_1)$  and  $\hat{\mathbb{P}}^{\text{out}} \leftarrow \hat{\mathbb{P}}_h$ 
15:       for step  $h = 1, 2, \dots, H$  do
16:        Take action  $a_h \leftarrow \pi_h(s_h)$  and observe next state  $s_{h+1}$ 
17:        Update  $N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$  and  $N_h(s_h, a_h, s_{h+1}) \leftarrow N_h(s_h, a_h, s_{h+1}) + 1$ 
18:         $\hat{\mathbb{P}}_h(\cdot | s_h, a_h) \leftarrow N_h(s_h, a_h, \cdot) / N_h(s_h, a_h)$ 
19: Return  $\hat{\mathbb{P}}^{\text{out}}$ 

```

number of times we have visited state-action (s, a) in step h before k^{th} episode in Algorithm 5. We use superscript k to denote variable corresponding to episode k ; in particular, $(s_1^k, a_1^k, \dots, s_H^k, a_H^k)$ is the trajectory we have visited in the k^{th} episode.

For any $\theta \in \mathcal{B}(1)$, let $\hat{\mathcal{M}}_\theta^k$ be the scalarized MDP using vector θ (defined in Section 2). We use $\hat{V}^k(\cdot; \theta)$, $\hat{Q}^k(\cdot, \cdot; \theta)$, and $\hat{\pi}_\theta^k = \hat{\pi}^k(\cdot; \theta)$ to denote the optimal value function, optimal Q-value function, and optimal policy of $\hat{\mathcal{M}}_\theta^k$ respectively. Therefore, we have

$$\begin{aligned}
\hat{Q}_h^k(s, a; \theta) &= [\hat{\mathbb{P}}_h^k \hat{V}_{h+1}^k](s, a; \theta) + \hat{r}_h^k(s, a; \theta), \\
\hat{V}_h^k(s; \theta) &= \max_{a \in \mathcal{A}} \hat{Q}_h^k(s, a; \theta), \\
\hat{\pi}_h^k(s; \theta) &= \operatorname{argmax}_{a \in \mathcal{A}} \hat{Q}_h^k(s, a; \theta).
\end{aligned} \tag{8}$$

Theorem 18. *There exist absolute constants c_β and c_K , such that for any $\epsilon \in (0, H]$, $\delta \in (0, 1]$, if we choose bonus $\beta_t = c_\beta(\sqrt{\min\{d, S\}H^2\iota/t} + H^2S\iota/t)$ where $\iota = \log[dSAKH/\delta]$, and run the exploration phase (Algorithm 5) for $K \geq c_K(\min\{d, S\}H^4SA\iota'/\epsilon^2 + H^3S^2A(\iota')^2/\epsilon)$ episodes where $\iota' = \log[dSAH/(\epsilon\delta)]$, then with probability at least $1 - \delta$, the algorithm satisfies*

$$\forall \theta \in \mathcal{B}(1) : \quad V_1^*(s_1; \theta) - V_1^{\pi_\theta}(s_1; \theta) \leq \epsilon,$$

where π_θ is the output of the any planning algorithm (e.g., value iteration) for the MDP $\hat{\mathcal{M}}_\theta^{\text{out}}$. Therefore, we have

$$m_{\text{RFE}}(\epsilon, \delta) \leq \mathcal{O}\left(\frac{\min\{d, S\}H^4SA\iota'}{\epsilon^2} + \frac{H^3S^2A(\iota')^2}{\epsilon}\right).$$

The bonus for episode k can be written as

$$\beta_h^k(s, a) = c_\beta \left(\sqrt{\frac{\min\{d, S\}H^2\iota}{\max\{N_h^k(s, a), 1\}}} + \frac{H^2S\iota}{\max\{N_h^k(s, a), 1\}} \right), \tag{9}$$

where $\iota = \log[dSAKH/\delta]$ and c_β is some large absolute constant.

We begin with the following lemma showing that the value function for a fixed π and also the optimal value function is H -Lipschitz with respect to θ .

Lemma 19. For all $(s, h) \in \mathcal{S} \times [H]$, for all policies π , and for any two vectors $\theta, \theta' \in \mathcal{B}(1)$, we have

$$\begin{aligned} |V_h^*(s; \theta) - V_h^*(s; \theta')| &\leq (H - h + 1) \|\theta - \theta'\| \\ |V_h^\pi(s; \theta) - V_h^\pi(s; \theta')| &\leq (H - h + 1) \|\theta - \theta'\| \end{aligned}$$

Proof of Lemma 19. We prove each claim separately.

First claim. We prove the lemma by backward induction on h . For $h = H + 1$ we have $V_h^*(s; \theta) = V_h^*(s; \theta') = 0$ and the inequality holds. Now assume that $|V_{h+1}^*(s; \theta) - V_{h+1}^*(s; \theta')| \leq (H - h) \|\theta - \theta'\|$ holds, we want to show that the claim also holds for h . We have

$$\begin{aligned} |V_h^*(s; \theta) - V_h^*(s; \theta')| &= \left| \max_{a \in \mathcal{A}} Q_h^*(s, a; \theta) - \max_{a' \in \mathcal{A}} Q_h^*(s, a'; \theta') \right| \\ &\leq \max_{a \in \mathcal{A}} |Q_h^*(s, a; \theta) - Q_h^*(s, a; \theta')| \\ &= \max_{a \in \mathcal{A}} |\langle \theta - \theta', \mathbf{r}_h(s, a) \rangle + \sum_{s' \in \mathcal{S}} \mathbb{P}(s' | s, a) (V_{h+1}^*(s'; \theta) - V_{h+1}^*(s'; \theta'))| \\ &\leq \max_{a \in \mathcal{A}} |\langle \theta - \theta', \mathbf{r}_h(s, a) \rangle| + \max_{a \in \mathcal{A}} \left| \sum_{s' \in \mathcal{S}} \mathbb{P}(s' | s, a) (V_{h+1}^*(s'; \theta) - V_{h+1}^*(s'; \theta')) \right| \\ &\leq \|\theta - \theta'\| + (H - h) \|\theta - \theta'\| \\ &= (H - h + 1) \|\theta - \theta'\|. \end{aligned}$$

It completes the proof of the lemma.

Second claim. The second claim is much easier to prove, since we have

$$\begin{aligned} |V_h^\pi(s; \theta) - V_h^\pi(s; \theta')| &= \left| \mathbb{E}_\pi \left[\sum_{h'=1}^H \langle \theta - \theta', \mathbf{r}_h(s'_h, a'_h) \rangle \right] \right| \\ &\leq \mathbb{E}_\pi \left[\sum_{h'=1}^H |\langle \theta - \theta', \mathbf{r}_h(s'_h, a'_h) \rangle| \right] \\ &\leq E_\pi \left[\sum_{h'=1}^H \|\theta - \theta'\| \right] \\ &= (H - h + 1) \|\theta - \theta'\| \end{aligned}$$

where the first inequality uses Jensen, and second inequality uses Cauchy-Schwarz. \square

Lemma 20. Let c be some large absolute constant such that $2c + 12c^2 \leq c_\beta$. Define event E_1 to be: for all $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$, $k \in [K]$, and $\theta \in \mathcal{B}(1)$,

$$\begin{cases} |[(\hat{\mathbb{P}}_h^k - \mathbb{P}_h) V_{h+1}^*](s, a; \theta)| &\leq c \sqrt{\frac{\min\{d, S\} H^2 \iota}{\max\{N_h^k(s, a), 1\}}}, \\ |(\hat{r}_h^k - r_h)(s, a; \theta)| &\leq c \sqrt{\frac{\iota}{\max\{N_h^k(s, a), 1\}}}, \\ |[(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)(s' | s, a)]| &\leq c \left(\sqrt{\frac{\hat{\mathbb{P}}_h^k(s' | s, a) \iota}{\max\{N_h^k(s, a), 1\}}} + \frac{\iota}{\max\{N_h^k(s, a), 1\}} \right), \end{cases} \quad (10)$$

where $\iota = \log[dSAKH/\delta]$. We have $\mathbb{P}(E_1) \geq 1 - \delta$.

Proof of Lemma 20. The proof is by applying concentration and covering arguments together with union bounds. The following shows that each claim holds with probability at least $1 - \delta$; rescaling δ to $\delta/3$ and applying a union bound completes the proof.

First claim: For a fixed $(s, a, k, h, \theta) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \times \mathcal{B}(1)$, using Azuma-Hoeffding inequality, with probability at least $1 - \delta'$ we have

$$|[(\hat{\mathbb{P}}_h^k - \mathbb{P}_h) V_{h+1}^*](s, a; \theta)| \leq \mathcal{O} \left(\sqrt{\frac{H^2 \log(1/\delta')}{N_h^k(s, a)}} \right).$$

Now consider an ϵ' -covering $\mathcal{B}_{\epsilon'}$ for the unit Euclidean ball $\mathcal{B}(1)$ with $\log |\mathcal{B}_{\epsilon'}| \leq \mathcal{O}(d \log(1/\epsilon'))$. For any $\boldsymbol{\theta} \in \mathcal{B}(1)$, there exists $\boldsymbol{\theta}' \in \mathcal{B}_{\epsilon'}$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \leq \epsilon'$. The concentration inequality above along with a union bound implies that with probability at least $1 - \delta$ for any $(s, a, k, h, \boldsymbol{\theta}') \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \times \mathcal{B}_{\epsilon'}$ we have

$$|[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^*](s, a; \boldsymbol{\theta}')| \leq \mathcal{O}\left(\sqrt{\frac{dH^2}{N_h^k(s, a)}} \log\left(\frac{SAKH}{\epsilon'\delta}\right)\right).$$

Now consider an arbitrary $(s, a, k, h, \boldsymbol{\theta}) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \times \mathcal{B}(1)$. Let $\boldsymbol{\theta}' \in \mathcal{B}_{\epsilon'}$ be such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \leq \epsilon'$; we have

$$\begin{aligned} & |[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^*](s, a; \boldsymbol{\theta})| \\ & \stackrel{(i)}{\leq} |[\widehat{\mathbb{P}}_h^k(V_{h+1}^*(\cdot; \boldsymbol{\theta}) - V_{h+1}^*(\cdot; \boldsymbol{\theta}'))](s, a)| + |[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^*](s, a; \boldsymbol{\theta}')| \\ & \quad + |[\mathbb{P}_h(V_{h+1}^*(\cdot; \boldsymbol{\theta}') - V_{h+1}^*(\cdot; \boldsymbol{\theta}))](s, a)| \\ & \stackrel{(ii)}{\leq} 2H\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| + \mathcal{O}\left(\sqrt{\frac{dH^2}{N_h^k(s, a)}} \log\left(\frac{SAKH}{\epsilon'\delta}\right)\right) \\ & \leq 2H\epsilon' + \mathcal{O}\left(\sqrt{\frac{dH^2}{N_h^k(s, a)}} \log\left(\frac{SAKH}{\epsilon'\delta}\right)\right), \end{aligned}$$

where (i) is by adding and subtracting the term $[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^*](s, a; \boldsymbol{\theta}')$ along with triangle inequality, and (ii) is by Lemma 19. Setting $\epsilon' = \frac{1}{HN_h^k(s, a)} \geq \frac{1}{HK}$ results in

$$|[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^*](s, a; \boldsymbol{\theta})| \leq \mathcal{O}\left(\sqrt{\frac{dH^2}{N_h^k(s, a)}} \log\left(\frac{SAKH}{\delta}\right)\right).$$

On the other hand, consider an ϵ' -cover $\mathcal{V}_{\epsilon'}$ for the ℓ_∞ ball of radius H in dimension S , i.e. $\{\mathbf{v} \in \mathbb{R}^S \mid \|\mathbf{v}\|_\infty \leq H\}$. For a fixed $(s, a, k, h, \mathbf{v}) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \times \mathcal{V}_{\epsilon'}$, using Azuma-Hoeffding inequality, with probability at least $1 - \delta'$ we have

$$|[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)\mathbf{v}](s, a)| \leq \mathcal{O}\left(\sqrt{\frac{H^2 \log(1/\delta')}{N_h^k(s, a)}}\right).$$

Note that $|\mathcal{V}_{\epsilon'}| \leq (3H/\epsilon')^d$, therefore by putting $\delta' = \delta/(SAKH|\mathcal{V}_{\epsilon'}|)$ we get for all $(s, a, k, h, \mathbf{v}) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \times \mathcal{V}_{\epsilon'}$

$$|[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)\mathbf{v}](s, a)| \leq \mathcal{O}\left(\sqrt{\frac{SH^2 \log(SAKH/(\epsilon'\delta))}{N_h^k(s, a)}}\right).$$

Now consider an arbitrary $(s, a, k, h, \boldsymbol{\theta}) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \times \mathcal{B}(1)$, and let $\mathbf{v} \in \mathcal{V}_{\epsilon'}$ be such that $\|V_{h+1}^*(\cdot; \boldsymbol{\theta}) - \mathbf{v}\|_\infty \leq \epsilon'$. We have

$$\begin{aligned} |[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^*](s, a; \boldsymbol{\theta})| & \leq |[\widehat{\mathbb{P}}_h^k(V_{h+1}^*(\cdot; \boldsymbol{\theta}) - \mathbf{v})](s, a)| + |[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)\mathbf{v}](s, a)| \\ & \quad + |[\mathbb{P}_h(V_{h+1}^*(\cdot; \boldsymbol{\theta}) - \mathbf{v})](s, a)| \\ & \leq 2\epsilon' + \mathcal{O}\left(\sqrt{\frac{SH^2 \log(SAKH/(\epsilon'\delta))}{N_h^k(s, a)}}\right). \end{aligned}$$

Setting $\epsilon' = \frac{1}{N_h^k(s, a)} \geq \frac{1}{K}$ results in

$$|[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^*](s, a; \boldsymbol{\theta})| \leq \mathcal{O}\left(\sqrt{\frac{SH^2}{N_h^k(s, a)}} \log\left(\frac{SAKH}{\delta}\right)\right)$$

The two bounds together complete the proof for the first claim.

Second claim: We have $\|\mathbf{r}_h^k\| \leq 1$ almost surely and $\mathbb{E}[\mathbf{r}_h^k \mid \mathcal{F}_h^k] = \mathbf{r}_h(s_h^k, a_h^k)$. For a fixed $(s, a, k, h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$, applying Lemma 36 implies that with probability at least $1 - \delta'$ we have

$$\|(\hat{\mathbf{r}}_h^k - \mathbf{r}_h)(s, a)\| \leq \mathcal{O}\left(\sqrt{\frac{\log(d/\delta')}{N_h^k(s, a)}}\right).$$

Setting $\delta' = \delta/(SAKH)$ and applying a union bound, for all $(s, a, k, h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$, we have

$$\|(\hat{\mathbf{r}}_h^k - \mathbf{r}_h)(s, a)\| \leq \mathcal{O}\left(\sqrt{\frac{\log(dSAKH/\delta)}{N_h^k(s, a)}}\right).$$

Now consider an arbitrary $(s, a, k, h, \boldsymbol{\theta}) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \times \mathcal{B}(1)$, we have (by Cauchy-Schwarz)

$$\begin{aligned} |(\hat{r}_h^k - r_h)(s, a; \boldsymbol{\theta})| &= |\langle \boldsymbol{\theta}, (\hat{\mathbf{r}}_h^k - \mathbf{r}_h)(s, a) \rangle| \\ &\leq \|\boldsymbol{\theta}\| \|(\hat{\mathbf{r}}_h^k - \mathbf{r}_h)(s, a)\| \\ &\leq \mathcal{O}\left(\sqrt{\frac{\log(dSAKH/\delta)}{N_h^k(s, a)}}\right), \end{aligned}$$

completing proof of this claim.

Third claim: For a fixed $(s, a, s', k, h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [K] \times [H]$, using empirical Bernstein inequality, with probability at least $1 - \delta'$ we have

$$|(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)(s' \mid s, a)| \leq \mathcal{O}\left(\sqrt{\frac{\hat{\mathbb{P}}_h^k(s' \mid s, a) \log(1/\delta')}{N_h^k(s, a)}} + \frac{\log(1/\delta')}{N_h^k(s, a)}\right)$$

Applying a union bound and setting $\delta' = \delta/S^2AKH$ completes the proof. \square

The following lemma shows that the optimal value functions of $\widehat{\mathcal{M}}_{\boldsymbol{\theta}}^k$ are close to the optimal value functions of $\mathcal{M}_{\boldsymbol{\theta}}$ and their difference is controlled by \tilde{Q} and \tilde{V} computed in Algorithm 5.

Lemma 21. Suppose event E_1 holds (defined in Lemma 20); then, for all $(s, a, k, h, \boldsymbol{\theta}) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \times \mathcal{B}(1)$ we have

$$\begin{aligned} |\hat{Q}_h^k(s, a; \boldsymbol{\theta}) - Q_h^*(s, a; \boldsymbol{\theta})| &\leq \tilde{Q}_h^k(s, a), \\ |\hat{V}_h^k(s; \boldsymbol{\theta}) - V_h^*(s; \boldsymbol{\theta})| &\leq \tilde{V}_h^k(s). \end{aligned} \tag{11}$$

Proof of Lemma 21. We prove the lemma by backward induction on h . For $h = H + 1$ the claim holds trivially. Now suppose that the claim is true for $(h + 1)^{\text{th}}$ step, we want to show that the claim is also true for h^{th} step. For the Q-value function we have

$$\begin{aligned} &|\hat{Q}_h^k(s, a; \boldsymbol{\theta}) - Q_h^*(s, a; \boldsymbol{\theta})| \\ &\leq \min \left\{ \underbrace{|[(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^*](s, a; \boldsymbol{\theta})|}_{(T_1)} + \underbrace{|[\hat{\mathbb{P}}_h^k(\hat{V}_{h+1}^k - V_{h+1}^*)](s, a; \boldsymbol{\theta})|}_{(T_2)}, H \right\} \\ &\stackrel{(i)}{\leq} \min \left\{ \beta_h^k(s, a) + [\hat{\mathbb{P}}_h^k \tilde{V}_{h+1}^k](s, a), H \right\} \stackrel{(ii)}{=} \tilde{Q}_h^k(s, a), \end{aligned}$$

where (i) follows from $T_1 \leq \beta_h^k(s, a)$ (event E_1) and $T_2 \leq [\hat{\mathbb{P}}_h^k \tilde{V}_{h+1}^k](s, a)$ (induction hypothesis), and (ii) is due to definition of \tilde{Q}_h^k in Algorithm 5. Now for the value function we have

$$\begin{aligned} &|\hat{V}_h^k(s; \boldsymbol{\theta}) - V_h^*(s; \boldsymbol{\theta})| \\ &= \left| \max_{a \in \mathcal{A}} \hat{Q}_h^k(s, a; \boldsymbol{\theta}) - \max_{a' \in \mathcal{A}} \hat{Q}^*(s, a'; \boldsymbol{\theta}) \right| \\ &\leq \max_{a \in \mathcal{A}} |\hat{Q}_h^k(s, a; \boldsymbol{\theta}) - \hat{Q}^*(s, a; \boldsymbol{\theta})| \\ &\leq \max_{a \in \mathcal{A}} \tilde{Q}_h^k(s, a) = \tilde{V}_h^k(s), \end{aligned}$$

which completes the induction step and consequently the proof. \square

Now we are ready to introduce the main lemma that shows value of $\hat{\pi}_\theta^k$ under the true model is close to its value under empirical model. The difference is controlled by \tilde{Q} and \tilde{V} computed in Algorithm 5.

Lemma 22. *Suppose event E_1 holds (defined in Lemma 20); then, for all $(s, a, k, h, \theta) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \times \mathcal{B}(1)$ we have*

$$\begin{aligned} |\hat{Q}_h^k(s, a; \theta) - Q_h^{\hat{\pi}_\theta^k}(s, a; \theta)| &\leq \alpha_h \tilde{Q}_h^k(s, a), \\ |\hat{V}_h^k(s; \theta) - V_h^{\hat{\pi}_\theta^k}(s; \theta)| &\leq \alpha_h \tilde{V}_h^k(s), \end{aligned} \quad (12)$$

where $\alpha_{H+1} = 1$ and $\alpha_h = [(1 + \frac{1}{H})\alpha_{h+1} + \frac{1}{H}]$; we have $1 \leq \alpha_h \leq 5$ for $h \in [H]$.

Proof of Lemma 22. We prove the claim by backward induction on h . For $h = H + 1$ the claim trivially holds. Now suppose that the claim is true for step $h + 1$ and we want to show that it also holds for step h .

$$\begin{aligned} &|\hat{Q}_h^k(s, a; \theta) - Q_h^{\hat{\pi}_\theta^k}(s, a; \theta)| \\ &\leq \min \left\{ \underbrace{|[(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)(V_{h+1}^{\hat{\pi}_\theta^k} - V_{h+1}^*)](s, a; \theta)|}_{(T_1)} \right. \\ &\quad + \underbrace{|[(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^*](s, a; \theta)| + |(\hat{r}_h^k - r_h)(s, a; \theta)|}_{(T_2)} \\ &\quad \left. + \underbrace{|[\hat{\mathbb{P}}_h^k(\hat{V}_{h+1}^k - V_{h+1}^{\hat{\pi}_\theta^k})](s, a; \theta)|, H}_{(T_3)} \right\} \end{aligned} \quad (13)$$

For the term (T_3) , by applying induction hypothesis we have

$$(T_3) \leq \alpha_{h+1} [\hat{\mathbb{P}}_h^k \tilde{V}_{h+1}^k](s, a). \quad (14)$$

Using event E_1 , for the term (T_2) we have

$$(T_2) \leq 2c \sqrt{\frac{\min\{d, S\} H^2 \iota}{\max\{N_h^k(s, a), 1\}}}. \quad (15)$$

It only remains to bound the term (T_1) ; we have

$$\begin{aligned} (T_1) &\leq \sum_{s' \in \mathcal{S}} |\hat{\mathbb{P}}_h^k(s' | s, a) - \mathbb{P}_h(s' | s, a)| |(V_{h+1}^{\hat{\pi}_\theta^k} - V_{h+1}^*)(s')| \\ &\leq \sum_{s' \in \mathcal{S}} |\hat{\mathbb{P}}_h^k(s' | s, a) - \mathbb{P}_h(s' | s, a)| \left[|(V_{h+1}^{\hat{\pi}_\theta^k} - \hat{V}_{h+1}^k)(s')| + |(\hat{V}_{h+1}^k - V_{h+1}^*)(s')| \right] \\ &\stackrel{(i)}{\leq} \sum_{s' \in \mathcal{S}} |\hat{\mathbb{P}}_h^k(s' | s, a) - \mathbb{P}_h(s' | s, a)| (\alpha_{h+1} + 1) \tilde{V}_{h+1}^k(s') \\ &\stackrel{(ii)}{\leq} \sum_{s' \in \mathcal{S}} \left[c \sqrt{\frac{\hat{\mathbb{P}}_h^k(s' | s, a) \iota}{\max\{N_h^k(s, a), 1\}}} + \frac{\iota}{\max\{N_h^k(s, a), 1\}} \right] (\alpha_{h+1} + 1) \tilde{V}_{h+1}^k(s') \\ &\stackrel{(iii)}{\leq} \sum_{s' \in \mathcal{S}} \left[\frac{\hat{\mathbb{P}}_h^k(s' | s, a)}{H} + \frac{c^2 H \iota + c \iota}{\max\{N_h^k(s, a), 1\}} \right] (\alpha_{h+1} + 1) \tilde{V}_{h+1}^k(s') \\ &\leq \frac{\alpha_{h+1} + 1}{H} [\hat{\mathbb{P}}_h^k \tilde{V}_{h+1}^k](s, a) + 2c^2 (\alpha_{h+1} + 1) \frac{H^2 S \iota}{\max\{N_h^k(s, a), 1\}}, \end{aligned} \quad (16)$$

where (i) is due Lemma 21 along with induction hypothesis, (ii) is due to event E_1 , and (iii) is by AM-GM. Plugging equation 14, 15, and 16 back in 13, we get

$$\begin{aligned}
& |\widehat{Q}_h^k(s, a; \theta) - \widetilde{Q}_h^k(s, a; \theta)| \\
& \leq \min \left\{ \left[\left(1 + \frac{1}{H}\right) \alpha_{h+1} + \frac{1}{h} \right] [\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k](s, a) + 2c \sqrt{\frac{\min\{d, S\} H^2 \iota}{\max\{N_h^k(s, a), 1\}}} \right. \\
& \quad \left. + 2c^2 (\alpha_{h+1} + 1) \frac{H^2 S \iota}{\max\{N_h^k(s, a), 1\}}, H \right\} \\
& \stackrel{(i)}{\leq} \min \left\{ \left[\left(1 + \frac{1}{H}\right) \alpha_{h+1} + \frac{1}{h} \right] [\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k](s, a) + \beta_h^k(s, a), H \right\} \\
& \stackrel{(ii)}{\leq} \alpha_h \min \{ [\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k](s, a) + \beta_h^k(s, a), H \} \\
& \stackrel{(iii)}{=} \alpha_h \widetilde{Q}_h^k(s, a),
\end{aligned} \tag{17}$$

where (i) is by the definition of the bonus β_h^k (we have $2c + 12c^2 \leq C$ and $(\alpha_{h+1} + 1) \leq 6$), (ii) is by the definition of α_h (note that $1 \leq \alpha_h$), and (iii) is by the definition of \widetilde{Q}_h^k in Algorithm 5. The inequality for value function follows immediately since we have

$$\begin{aligned}
& |\widehat{V}_h^k(s; \theta) - V_h^k(s; \theta)| \\
& = |[\mathbb{D}_{\widehat{\pi}_\theta^k} \widehat{Q}_h^k](s; \theta) - [\mathbb{D}_{\pi_\theta^k} Q_h^k](s; \theta)| \\
& \leq \alpha_h [\mathbb{D}_{\pi_\theta^k} \widetilde{Q}_h^k](s) \\
& \leq \alpha_h \max_{a \in \mathcal{A}} \widetilde{Q}_h^k(s, a) \\
& = \alpha_h \widetilde{V}_h^k(s).
\end{aligned}$$

It completes the induction step and consequently the proof of the lemma. \square

Theorem 23 (Similar to guarantee for UCB-VI from Azar et al. 2017). *For any $\delta \in (0, 1]$, if we choose β_t^k in Algorithm 5 as in Equation 9; then, with probability at least $1 - \delta$, we have*

$$\sum_{k=1}^K \widetilde{V}_1^k(s_1) \leq \mathcal{O}(\sqrt{\min\{d, S\} H^4 S A K \iota} + H^3 S^2 A \iota^2).$$

Proof of Theorem 23. For a fixed k , by definition of \widetilde{V} we have

$$\widetilde{V}_1^k(s_1) \leq \sum_{h=1}^H (\beta_h^k(s_h^k, a_h^k) + \zeta_h^k),$$

where $\zeta_h^k = [\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k](s_h^k, a_h^k) - \widetilde{V}_{h+1}^k(s_{h+1}^k)$. Summing over k gives us,

$$\sum_{k=1}^K \widetilde{V}_1^k(s_1) \leq \underbrace{\sum_{k=1}^K \sum_{h=1}^H \beta_h^k(s_h^k, a_h^k)}_{(T_1)} + \underbrace{\sum_{k=1}^K \sum_{h=1}^H \zeta_h^k}_{(T_2)}.$$

Now we bound each term separately. For the term (T_1) , using standard pigeonhole argument, we have

$$\begin{aligned}
(T_1) &= C \left[\sum_{k=1}^K \sum_{h=1}^H \sqrt{\frac{\min\{d, S\} H^2 \iota}{N_h^k(s_h^k, a_h^k)}} + \sum_{k=1}^K \sum_{h=1}^H \frac{H^2 S \iota}{N_h^k(s_h^k, a_h^k)} \right] \\
&= C \left[\sqrt{\min\{d, S\} H^2 \iota} \sum_{h,s,a} \sum_{i=1}^{N_h^K(s,a)} \sqrt{\frac{1}{i}} + H^2 S \iota \sum_{h,s,a} \sum_{i=1}^{N_h^K(s,a)} \frac{1}{i} \right] \\
&\leq C' \left[\sqrt{\min\{d, S\} H^2 \iota} \sum_{h,s,a} \sqrt{N_h^K(s,a)} + H^2 S \iota \sum_{h,s,a} \log(KH) \right] \\
&\leq C' \left[\sqrt{\min\{d, S\} H^2 \iota} \sqrt{HSA} \sqrt{KH} + H^3 S^2 A \iota^2 \right] \\
&\leq \mathcal{O}(\sqrt{\min\{d, S\} H^4 SAK \iota} + H^3 S^2 A \iota^2).
\end{aligned}$$

For the second term, note that ζ_h^k forms a martingale difference sequence; therefore, by Azuma-Hoeffding's inequality, with probability at least $1 - \delta$, we have

$$(T_2) \leq \mathcal{O}(H \sqrt{(KH) \log(1/\delta)}) = \mathcal{O}(\sqrt{H^3 K \log(1/\delta)}),$$

resulting in a lower order term and completing the proof. \square

Proof of Theorem 18. By Algorithm 5, we have $\text{out} = \arg\min_{k \in [K]} \tilde{V}_1^k(s_1)$, resulting in $\tilde{V}_1^{\text{out}}(s_1) \leq \frac{1}{K} \sum_{k=1}^K \tilde{V}_1^k(s_1)$. Therefore, with probability at least $1 - 2\delta$, for any vector $\theta \in \mathcal{B}(1)$ we have

$$\begin{aligned}
V_1^*(s_1; \theta) - V_1^{\hat{\pi}_{\theta}^{\text{out}}}(s_1; \theta) &\leq |V_1^*(s_1; \theta) - \hat{V}_1^{\text{out}}(s_1; \theta)| + |\hat{V}_1^{\text{out}}(s_1; \theta) - V_1^{\hat{\pi}_{\theta}^{\text{out}}}(s_1; \theta)| \\
&\stackrel{(i)}{\leq} (1 + \alpha_1) \tilde{V}_1^{\text{out}}(s_1) \\
&\leq 6 \tilde{V}_1^{\text{out}}(s_1) \\
&\leq \frac{6}{K} \sum_{k=1}^K \tilde{V}_1^k(s_1) \\
&\stackrel{(ii)}{\leq} \mathcal{O}(\sqrt{\min\{d, S\} H^4 S A \iota / K} + H^3 S^2 A \iota^2 / K) \\
&\stackrel{(iii)}{\leq} \epsilon,
\end{aligned}$$

where (i) is due to Lemma 21 and Lemma 22, (ii) is due to Theorem 23, and (iii) is due to $K \geq c_K (\min\{d, S\} H^4 S A \iota' / \epsilon^2 + H^3 S^2 A (\iota')^2 / \epsilon)$ with a sufficiently large constant c_K . Rescaling δ completes the proof. \square

D PROOF FOR SECTION 5

In this section we provide proofs and missing details for Section 5.

D.1 REWARD-FREE ALGORITHM FOR LINEAR VMDPS

We use slightly modified version of the reward-free algorithm introduced by Wang et al. (2020). The exploration phase and planning phase are displayed in Algorithm 6 and 7, respectively.

D.2 PROOF OF THEOREM 9

In this section, we prove Theorem 24 which implies the first claim in Theorem 9. Second and third claims in Theorem 9 immediately follow due to Theorem 6 and Theorem 5.

Theorem 24. *There exist absolute constants c_β and c_K , such that for any $\epsilon \in (0, H]$ and $\delta \in (0, 1]$, if we choose bonus coefficient $\beta = c_\beta \cdot d_{\text{lin}} H \sqrt{\iota}$ with $\iota = \log[d_{\text{lin}} d K H / \delta]$, and run the exploration*

Algorithm 6 Reward-Free RL for Linear VMDPs: Exploration Phase

```

1: Hyperparameters: Bonus coefficient  $\beta$ .
2: for episode  $k = 1, 2, \dots, K$  do
3:   for step  $h = H, H-1, \dots, 1$  do
4:      $\tilde{\Lambda}_h^k = \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i) \phi(s_h^i, a_h^i)^\top + I$ 
5:      $\tilde{u}_h^k(\cdot, \cdot) \leftarrow \min\{\beta \cdot \sqrt{\phi(\cdot, \cdot)^\top (\tilde{\Lambda}_h^k)^{-1} \phi(\cdot, \cdot)}, H\}$ 
6:     Define  $\tilde{r}_h^k(\cdot, \cdot) \leftarrow \tilde{u}_h^k(\cdot, \cdot)/H$ 
7:      $\tilde{\mathbf{w}}_h^k \leftarrow (\tilde{\Lambda}_h^k)^{-1} \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i) \tilde{V}_{h+1}^k(s_{h+1}^i)$ 
8:      $\tilde{Q}_h^k(\cdot, \cdot) \leftarrow \min\{(\tilde{\mathbf{w}}_h^k)^\top \phi(\cdot, \cdot) + \tilde{r}_h^k(\cdot, \cdot) + \tilde{u}_h^k(\cdot, \cdot), H\}$ 
9:      $\tilde{V}_h^k(\cdot) = \max_{a \in \mathcal{A}} \tilde{Q}_h^k(\cdot, a)$  and  $\tilde{\pi}_h^k(\cdot) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \tilde{Q}_h^k(\cdot, a)$ 
10:    Observe initial state  $s_1^k \leftarrow s_1$ 
11:    for step  $h = 1, 2, \dots, H$  do
12:      Take action  $a_h^k \leftarrow \tilde{\pi}_h^k(s_h^k)$  and observe next state  $s_{h+1}^k$ 
13: Return  $\mathcal{D} \leftarrow \{(s_h^k, a_h^k)\}_{(h,k) \in [H] \times [K]}$ 

```

Algorithm 7 Reward-Free RL for Linear VMDPs: Planning Phase

```

1: Hyperparameters: Bonus coefficient  $\beta$ .
2: Input: Dataset  $\mathcal{D} = \{(s_h^k, a_h^k)\}_{(k,h) \in [K] \times [H]}$ , vector  $\boldsymbol{\theta} \in \mathcal{B}(1)$ 
   samples of return function  $\{\mathbf{r}_h^k\}_{(k,h) \in [K] \times [H]}$ 
3: for step  $h = H, H-1, \dots, 1$  do
4:    $\hat{\Lambda}_h = \sum_{i=1}^K \phi(s_h^i, a_h^i) \phi(s_h^i, a_h^i)^\top + I$ 
5:    $\hat{u}_h(\cdot, \cdot) \leftarrow \min\{\beta \cdot \sqrt{\phi(\cdot, \cdot)^\top (\hat{\Lambda}_h)^{-1} \phi(\cdot, \cdot)}, H\}$ 
6:    $\hat{\mathbf{w}}_h \leftarrow (\hat{\Lambda}_h)^{-1} \sum_{i=1}^K \phi(s_h^i, a_h^i) [\hat{V}_{h+1}(s_{h+1}^i) + \boldsymbol{\theta}^\top \mathbf{r}_h^i]$ 
7:    $\hat{Q}_h(\cdot, \cdot) \leftarrow \min\{(\hat{\mathbf{w}}_h)^\top \phi(\cdot, \cdot) + \hat{u}_h(\cdot, \cdot), H\}$ 
8:    $\hat{V}_h(\cdot) = \max_{a \in \mathcal{A}} \hat{Q}_h(\cdot, a)$  and  $\hat{\pi}_h(\cdot) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \hat{Q}_h(\cdot, a)$ 
9: Return  $\pi_\theta = \{\hat{\pi}_h\}_{h=1}^H$ 

```

algorithm (Algorithm 6) for $K \geq c_K [d_{\text{lin}}^3 H^6 (\iota')^2 / \epsilon^2]$ episodes where $\iota' = \log[d_{\text{lin}} dH / (\epsilon \delta)]$, then with probability at least $1 - \delta$, for any $\theta \in \mathcal{B}(1)$, the output of the planning phase satisfies:

$$V_1^*(s_1; \theta) - V_1^{\pi_\theta}(s_1; \theta) \leq \epsilon,$$

where π_θ is the output of the planning algorithm (Algorithm 7) given θ as input. Therefore, in this case we have

$$m_{\text{RFE}}(\epsilon, \delta) \leq \mathcal{O}\left(d_{\text{lin}}^3 H^6 (\iota')^2 / \epsilon^2\right).$$

In this section, we denote $\phi_h^k := \phi(s_h^k, a_h^k)$ for $(k, h) \in [K] \times [H]$. For a scalar reward function $r' : \mathcal{S} \times \mathcal{A} \rightarrow [-1, 1]$ and a policy π , we use $V_h^\pi(\cdot | r')$ and $Q_h^\pi(\cdot, \cdot | r')$ to denote the value function and Q-value function for the MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r')$. Similarly we define the optimal value function and Q-value function denoted by $V_h^*(\cdot | r')$ and $Q_h^*(\cdot, \cdot | r')$.

The bonus coefficient is defined to be

$$\beta = c_\beta \cdot d_{\text{lin}} H \sqrt{\iota} \quad (18)$$

where $\iota = \log[d_{\text{lin}} dHK / \delta]$.

We start with the following concentration lemma.

Lemma 25. Suppose Assumption 8 holds. Let c be some large absolute constant. Define event E_2 to be: for all $(k, h, \theta) \in [K] \times [H] \times \mathcal{B}(1)$,

$$\begin{cases} \left\| \sum_{i=1}^{k-1} \phi_h^i \left(\tilde{V}_{h+1}^k(s_{h+1}^i) - [\mathbb{P}_h \tilde{V}_{h+1}^k](s_h^i, a_h^i) \right) \right\|_{(\tilde{\Lambda}_h^k)^{-1}} & \leq c \cdot H \sqrt{d_{\text{lin}}^2 \iota}, \\ \left\| \sum_{i=1}^K \phi_h^i \left(\hat{V}_{h+1}(s_{h+1}^i) - [\mathbb{P}_h \hat{V}_{h+1}](s_h^i, a_h^i) \right) \right\|_{(\hat{\Lambda}_h)^{-1}} & \leq c \cdot H \sqrt{d_{\text{lin}}^2 \iota}, \\ \left\| \sum_{i=1}^K \phi_h^i \left(\theta^\top (\hat{\mathbf{r}}_h - \mathbf{r}_h)(s_h^i, a_h^i) \right) \right\|_{(\hat{\Lambda}_h)^{-1}} & \leq c \cdot \sqrt{d_{\text{lin}} \iota}, \\ \left| \sum_{k=1}^K \sum_{h=1}^H [\mathbb{P}_h \tilde{V}_{h+1}^k](s_h^k, a_h^k) - \tilde{V}_{h+1}^k(s_h^k) \right| & \leq c \cdot H^2 \sqrt{K \iota}, \end{cases} \quad (19)$$

where $\iota = \log[d_{\text{lin}} dHK / \delta]$. We have $\mathbb{P}(E_2) \geq 1 - \delta$.

Proof of Lemma 25. The first three inequalities follow from the standard concentration inequalities of the self-normalized process, a covering argument over the value functions or θ , and union bound. We refer readers to the proofs of Lemma B.3 in Jin et al. (2020b) or Lemma A.1 in Wang et al. (2020) for details. The last inequality follows immediately from Azuma-Hoeffding's inequality since for a fixed h , $\{\mathbb{P}_h \tilde{V}_{h+1}^k(s_h^k, a_h^k) - \tilde{V}_{h+1}^k(s_h^k)\}_{k \in [K]}$ is a martingale difference sequence bounded by H . \square

The following lemma shows that \tilde{V}_1^k (defined in Algorithm 6) is optimistic with respect to reward function \tilde{r}^k . In addition, it shows its sum over k can be controlled by $\tilde{\mathcal{O}}(\sqrt{d_{\text{lin}}^3 H^4 K})$.

Lemma 26. Suppose Assumption 8 and event E_2 (defined in Lemma 25) hold; we have

$$\begin{aligned} V_1^*(s_1 | \tilde{r}^k) &\leq \tilde{V}_1^k(s_1) \quad \forall k \in [K] \\ \sum_{k=1}^K \tilde{V}_1^k(s_1) &\leq \mathcal{O}\left(\sqrt{d_{\text{lin}}^3 H^4 K \iota^2}\right) \end{aligned}$$

Proof of Lemma 26. Let $\bar{\mathbf{w}}_h^k = \int \tilde{V}_{h+1}^k(s') d\mu_h(s')$; by Assumption 8, we have

$$\begin{aligned} \|\bar{\mathbf{w}}_h^k\| &\leq H \|\mu_h(\mathcal{S})\| \leq H \sqrt{d_{\text{lin}}} \\ [\mathbb{P}_h \tilde{V}_{h+1}^k](s, a) &= \phi(s, a)^\top \bar{\mathbf{w}}_h^k \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A} \end{aligned} \quad (20)$$

For all $k, h, s, a \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}
& \phi(s, a)^\top \tilde{\mathbf{w}}_h^k - [\mathbb{P}_h \tilde{V}_{h+1}^k](s, a) \\
&= \phi(s, a)^\top [\tilde{\mathbf{w}}_h^k - \bar{\mathbf{w}}_h^k] \\
&= \phi(s, a)^\top (\tilde{\Lambda}_h^k)^{-1} \left(\sum_{i=1}^{k-1} \phi_h^i \tilde{V}_{h+1}^k(s_{h+1}^i) - \tilde{\Lambda}_h^k \bar{\mathbf{w}}_h^k \right) \\
&= \phi(s, a)^\top (\tilde{\Lambda}_h^k)^{-1} \left(\sum_{i=1}^{k-1} \phi_h^i \tilde{V}_{h+1}^k(s_{h+1}^i) - \sum_{i=1}^{k-1} \phi_h^i (\phi_h^i)^\top \bar{\mathbf{w}}_h^k - \bar{\mathbf{w}}_h^k \right).
\end{aligned}$$

Note that $(\phi_h^i)^\top \bar{\mathbf{w}}_h^k = [\mathbb{P}_h \tilde{V}_{h+1}^k](s_h^i, a_h^i)$. Therefore, we have

$$\begin{aligned}
& |\phi(s, a)^\top \tilde{\mathbf{w}}_h^k - [\mathbb{P}_h \tilde{V}_{h+1}^k](s, a)| \\
&= \left| \phi(s, a)^\top (\tilde{\Lambda}_h^k)^{-1} \left[\sum_{i=1}^{k-1} \phi_h^i \left(\tilde{V}_{h+1}^k(s_{h+1}^i) - [\mathbb{P}_h \tilde{V}_{h+1}^k](s_h^i, a_h^i) \right) - \bar{\mathbf{w}}_h^k \right] \right| \\
&\leq \left| \phi(s, a)^\top (\tilde{\Lambda}_h^k)^{-1} \left[\sum_{i=1}^{k-1} \phi_h^i \left(\tilde{V}_{h+1}^k(s_{h+1}^i) - [\mathbb{P}_h \tilde{V}_{h+1}^k](s_h^i, a_h^i) \right) \right] \right| + |\phi(s, a)^\top (\tilde{\Lambda}_h^k)^{-1} \bar{\mathbf{w}}_h^k| \\
&\leq \left\| \sum_{i=1}^{k-1} \phi_h^i \left(\tilde{V}_{h+1}^k(s_{h+1}^i) - [\mathbb{P}_h \tilde{V}_{h+1}^k](s_h^i, a_h^i) \right) \right\|_{(\tilde{\Lambda}_h^k)^{-1}} \cdot \|\phi(s, a)\|_{(\tilde{\Lambda}_h^k)^{-1}} + \|\bar{\mathbf{w}}_h^k\|_{(\tilde{\Lambda}_h^k)^{-1}} \cdot \|\phi(s, a)\|_{(\tilde{\Lambda}_h^k)^{-1}}.
\end{aligned}$$

Note that $\|\bar{\mathbf{w}}_h^k\|_{(\tilde{\Lambda}_h^k)^{-1}} \leq \|\bar{\mathbf{w}}_h^k\| \leq H\sqrt{d_{\text{lin}}}$ since $\tilde{\Lambda}_h^k \succeq I$. By event E_2 we have

$\left\| \sum_{i=1}^{k-1} \phi_h^i \left(\tilde{V}_{h+1}^k(s_{h+1}^i) - [\mathbb{P}_h \tilde{V}_{h+1}^k](s_h^i, a_h^i) \right) \right\|_{(\tilde{\Lambda}_h^k)^{-1}} \leq c \cdot H\sqrt{d_{\text{lin}}^2 \iota}$. Plugging back, results in

$$\begin{aligned}
& |\phi(s, a)^\top \tilde{\mathbf{w}}_h^k - [\mathbb{P}_h \tilde{V}_{h+1}^k](s, a)| \\
&\leq (H\sqrt{d_{\text{lin}}} + c \cdot H\sqrt{d_{\text{lin}}^2 \iota}) \|\phi(s, a)\|_{(\tilde{\Lambda}_h^k)^{-1}} \\
&\leq (c_\beta \cdot H\sqrt{d_{\text{lin}}^2 \iota}) \|\phi(s, a)\|_{(\tilde{\Lambda}_h^k)^{-1}} \\
&= \beta \|\phi(s, a)\|_{(\tilde{\Lambda}_h^k)^{-1}}
\end{aligned} \tag{21}$$

Now we are ready to complete the proof:

First claim: we prove the claim

$$V_h^*(s \mid \tilde{r}^k) \leq \tilde{V}_h^k(s) \quad \forall s \in \mathcal{S},$$

by backward induction on h . For $h = H + 1$ the claim is trivial since both LHS and RHS are zero. Now suppose that we have

$$V_{h+1}^*(s \mid \tilde{r}^k) \leq \tilde{V}_{h+1}^k(s) \quad \forall s \in \mathcal{S}.$$

Then, for all $s \in \mathcal{S}$ we have

$$\begin{aligned}
V_h^*(s \mid \tilde{r}^k) &= \max_{a \in \mathcal{A}} Q_h^*(s \mid \tilde{r}^k) \\
&= \max_{a \in \mathcal{A}} \{ \min \{ \tilde{r}_h^k(s, a) + [\mathbb{P}_h V_{h+1}^*](s, a \mid \tilde{r}^k), H \} \} \\
&\leq \max_{a \in \mathcal{A}} \{ \min \{ \tilde{r}_h^k(s, a) + [\mathbb{P}_h \tilde{V}_{h+1}^k](s, a), H \} \} \\
&\leq \max_{a \in \mathcal{A}} \{ \min \{ \tilde{r}_h^k(s, a) + \phi(s, a)^\top \tilde{\mathbf{w}}_h^k + \beta \|\phi(s, a)\|_{(\tilde{\Lambda}_h^k)^{-1}}, H \} \} \\
&\leq \max_{a \in \mathcal{A}} \tilde{Q}_h^k(s, a) = \tilde{V}_h^k(s),
\end{aligned}$$

where the first inequality is due to induction hypothesis and the second inequality is due to Equation 21. It proves the induction step and completes the induction.

Second claim: Let

$$\zeta_h^k = [\mathbb{P}_h \tilde{V}_{h+1}^k](s_h^k, a_h^k) - \tilde{V}_{h+1}^k(s_h^k) \quad \forall (k, h) \in [K] \times [H]$$

we have

$$\begin{aligned} \sum_{k=1}^K \tilde{V}_1^k(s_1^k) &\leq \sum_{k=1}^K ((\tilde{r}_1^k + u_1^k)(s_1^k, a_1^k) + (\phi_1^k)^\top \tilde{\mathbf{w}}_1^k) \\ &= \sum_{k=1}^K ((1 + 1/H)\beta \cdot \|\phi(s, a)\|_{(\tilde{\Lambda}_1^k)^{-1}} + (\phi_1^k)^\top \tilde{\mathbf{w}}_1^k) \\ &\leq \sum_{k=1}^K ((2 + 1/H)\beta \cdot \|\phi(s, a)\|_{(\tilde{\Lambda}_1^k)^{-1}} + [\mathbb{P}_1 \tilde{V}_2^k][s_1^k, a_1^k]) \\ &\leq \sum_{k=1}^K (\tilde{V}_2^k(s_2^k) + (2 + 1/H)\beta \cdot \|\phi(s, a)\|_{(\tilde{\Lambda}_1^k)^{-1}} + \zeta_1^k) \end{aligned}$$

By repeatedly applying the same argument we get

$$\sum_{k=1}^K \tilde{V}_1^k(s_1^k) \leq (2 + 1/H)\beta \underbrace{\sum_{k=1}^K \sum_{h=1}^H \|\phi(s, a)\|_{(\tilde{\Lambda}_h^k)^{-1}}}_{(T_1)} + \underbrace{\sum_{k=1}^K \sum_{h=1}^H \zeta_h^k}_{(T_2)}.$$

For the term (T_1) we have

$$\begin{aligned} T_1 &= \sum_{k=1}^K \sum_{h=1}^H \|\phi(s, a)\|_{(\tilde{\Lambda}_h^k)^{-1}} \\ &\stackrel{(i)}{\leq} \sqrt{KH \sum_{k=1}^K \sum_{h=1}^H (\phi_h^k)^\top (\tilde{\Lambda}_h^k) (\phi_h^k)} \\ &\stackrel{(ii)}{\leq} \sqrt{KH(2d_{\text{lin}}H \log(K))}, \end{aligned}$$

where (i) uses Cauchy-Schwarz, and (ii) uses Lemma D.2 in Jin et al. (2020b) that implies $\sum_{k=1}^K \sum_{h=1}^H (\phi_h^k)^\top (\tilde{\Lambda}_h^k) (\phi_h^k) \leq 2d_{\text{lin}}H \log(K)$.

For the term (T_2) , by the third inequality in event E_2 , we have

$$T_2 \leq c \cdot H^2 \sqrt{K\iota}.$$

Plugging back in the original equation gives us

$$\begin{aligned} \sum_{k=1}^K \tilde{V}_1^k(s_1^k) &\leq (2 + 1/H)\beta \cdot \sqrt{KH(2d_{\text{lin}}H \log(K))} + c \cdot H^2 \sqrt{K\iota} \\ &\leq c' \sqrt{d_{\text{lin}}^3 H^4 K \iota^2}, \end{aligned}$$

for some absolute constant c' , which completes the proof of the lemma. \square

Lemma 27. Suppose Assumption 8 and event E_2 (defined in Lemma 25) hold; Let $\hat{u} = \{\hat{u}_h\}_{h=1}^H$ (as defined in Line 5 of Algorithm 7), we have

$$V_1^*(s_1 \mid \hat{u}/H) \leq \mathcal{O}\left(\sqrt{d_{\text{lin}}^3 H^4 \iota^2 / K}\right)$$

Proof of Lemma 27. Note that $\hat{\Lambda}_h \succeq \tilde{\Lambda}_h^k$ for all $k \in [K]$. Therefore for all $h \in [H]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\hat{u}_h(s, a)/H \leq \tilde{u}_h^k(s, a)/H = \tilde{r}_h^k(s, a)$$

Using Lemma 26 we have

$$\begin{aligned} KV_1^*(s_1 \mid \hat{u}/H) &\leq \sum_{k=1}^K V_1^*(s_1 \mid \tilde{r}^k) \\ &\leq \sum_{k=1}^K \tilde{V}_1^k(s_1) \\ &\leq \mathcal{O}\left(\sqrt{d_{\text{lin}}^3 H^4 K \iota^2}\right). \end{aligned}$$

Dividing both sides by K completes the proof. \square

Lemma 28. Suppose Assumption 8 and event E_2 (defined in Lemma 25) hold. For all $(s, a, h, \theta) \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{B}(1)$ we have

$$Q_h^*(s, a; \theta) \leq \hat{Q}_h(s, a) \leq \theta^\top \mathbf{r}_h(s, a) + [\mathbb{P}_h \hat{V}_{h+1}](s, a) + 2\hat{u}_h(s, a).$$

Proof of Lemma 28. First note that by Assumption 8, we have $\mathbf{r}_h(s, a) = W_h \phi(s, a)$. Define

$$\bar{\mathbf{w}}_h = \int \hat{V}_{h+1}(s') d\mu_h(s') + \theta^\top W_h.$$

By Assumption 8, we have

$$\begin{aligned} \|\bar{\mathbf{w}}_h\| &\leq \left\| \int \hat{V}_{h+1}(s') d\mu_h(s') \right\| + \|\theta^\top W_h\| \\ &\leq H \|\mu_h(\mathcal{S})\| + \|\theta\| \|W_h\| \\ &\leq H \cdot \sqrt{d_{\text{lin}}} + \sqrt{d_{\text{lin}}} \leq 2H \sqrt{d_{\text{lin}}}. \end{aligned}$$

Therefore we have

$$\begin{aligned} \|\bar{\mathbf{w}}_h\| &\leq 2H \cdot \sqrt{d_{\text{lin}}} \\ [\mathbb{P}_h \hat{V}_{h+1}](s, a) + \theta^\top \mathbf{r}_h(s, a) &= \phi(s, a)^\top \bar{\mathbf{w}}_h \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A} \end{aligned} \tag{22}$$

Now using similar argument in Lemma 26, for all $(s, a, h, \theta) \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{B}(1)$ we can have

$$\begin{aligned} &|\phi(s, a)^\top \hat{\mathbf{w}}_h - [\mathbb{P}_h \hat{V}_{h+1}](s, a) - \theta^\top \mathbf{r}_h(s, a)| \\ &\leq \underbrace{\left\| \sum_{i=1}^K \phi_h^i \left(\hat{V}_{h+1}(s_{h+1}^i) - [\mathbb{P}_h \hat{V}_{h+1}](s_h^i, a_h^i) \right) \right\|_{(\hat{\Lambda}_h)^{-1}}}_{(T_1)} \cdot \|\phi(s, a)\|_{(\hat{\Lambda}_h)^{-1}} \\ &\quad + \underbrace{\left\| \sum_{i=1}^K \phi_h^i \left(\theta^\top (\hat{\mathbf{r}}_h - \mathbf{r}_h)(s_h^i, a_h^i) \right) \right\|_{(\hat{\Lambda}_h)^{-1}}}_{(T_2)} \cdot \|\phi(s, a)\|_{(\hat{\Lambda}_h)^{-1}} \\ &\quad + \underbrace{\|\bar{\mathbf{w}}_h^k\|_{(\hat{\Lambda}_h)^{-1}}}_{(T_3)} \cdot \|\phi(s, a)\|_{(\hat{\Lambda}_h)^{-1}} \end{aligned}$$

Note that $(T_3) = \|\bar{\mathbf{w}}_h\|_{(\hat{\Lambda}_h)^{-1}} \leq \|\bar{\mathbf{w}}_h\| \leq 2H \sqrt{d_{\text{lin}}}$ since $\hat{\Lambda}_h \succeq I$. The other two terms (T_1) and (T_2) are both upper-bounded by $c \cdot H \sqrt{d_{\text{lin}}^2 \iota}$ due to event E_2 . Plugging back results in

$$\begin{aligned} &|\phi(s, a)^\top \hat{\mathbf{w}}_h - [\mathbb{P}_h \hat{V}_{h+1}](s, a) - \theta^\top \mathbf{r}_h(s, a)| \\ &\leq \left[2H \sqrt{d_{\text{lin}}} + 2cH \sqrt{d_{\text{lin}}^2 \iota} \right] \|\phi(s, a)\|_{(\hat{\Lambda}_h)^{-1}} \\ &\leq [c_\beta \cdot H \sqrt{d_{\text{lin}}^2 \iota}] \|\phi(s, a)\|_{(\hat{\Lambda}_h)^{-1}} \\ &= \beta \|\phi(s, a)\|_{(\hat{\Lambda}_h)^{-1}}. \end{aligned} \tag{23}$$

Now we are ready to complete the proof of the lemma. For all $(s, a, h, \theta) \leq \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{B}(1)$, we have

$$\begin{aligned}\widehat{Q}_h(s, a) &= \min\{\phi(s, a)^\top \widehat{\mathbf{w}}_h + \widehat{u}_h(s, a), H\} \\ &\leq \min\{[\mathbb{P}_h \widehat{V}_{h+1}](s, a) + \theta^\top \mathbf{r}_h(s, a) + 2\beta \|\phi(s, a)\|_{(\widehat{\Lambda}_h)^{-1}}, H\} \\ &\leq [\mathbb{P}_h \widehat{V}_{h+1}](s, a) + \theta^\top \mathbf{r}_h(s, a) + 2\min\{\beta \|\phi(s, a)\|_{(\widehat{\Lambda}_h)^{-1}}, H\} \\ &= [\mathbb{P}_h \widehat{V}_{h+1}](s, a) + \theta^\top \mathbf{r}_h(s, a) + 2\widehat{u}_h(s, a),\end{aligned}$$

where the first inequality uses Equation 23. It completes the proof for one side of the inequality in Lemma 28. For the other side we prove the claim by backward induction on h . For $h = H + 1$ we the claim is trivial. Now suppose that

$$Q_{h+1}^*(s, a; \theta) \leq \widehat{Q}_{h+1}(s, a),$$

we want to prove the claim for h . We have

$$\begin{aligned}Q_h^*(s, a; \theta) &= \min\{\theta^\top \mathbf{r}_h(s, a) + [\mathbb{P}_h V_{h+1}^*](s, a; \theta), H\} \\ &\stackrel{(i)}{\leq} \min\{\theta^\top \mathbf{r}_h(s, a) + [\mathbb{P}_h \widehat{V}_{h+1}](s, a; \theta), H\} \\ &\stackrel{(ii)}{\leq} \min\{\phi(s, a)^\top \widehat{\mathbf{w}}_h + \beta \|\phi(s, a)\|_{(\widehat{\Lambda}_h)^{-1}}, H\} \\ &\leq \min\{\phi(s, a)^\top \widehat{\mathbf{w}}_h + \min\{\beta \|\phi(s, a)\|_{(\widehat{\Lambda}_h)^{-1}}, H\}, H\} \\ &= \min\{\phi(s, a)^\top \widehat{\mathbf{w}}_h + \widehat{u}_h(s, a), H\} = \widehat{Q}_h(s, a),\end{aligned}$$

where (i) uses induction hypothesis, and (ii) uses Equation 23. It completes the proof of the lemma. \square

Proof of Theorem 24. With probability at least $1 - \delta$, event E_2 holds and we have

$$\begin{aligned}\widehat{V}_1(s_1) - V_1^{\widehat{\pi}}(s_1; \theta) &= \widehat{Q}_1(s_1, \widehat{\pi}_1(s_1)) - Q_1^{\widehat{\pi}}(s_1, \widehat{\pi}_1(s_1); \theta) \\ &\stackrel{(i)}{\leq} \left([\mathbb{P}_1 \widehat{V}_2](s_1, \widehat{\pi}_1(s_1)) + \theta^\top \mathbf{r}_1(s_1, \widehat{\pi}_1(s_1)) + 2\widehat{u}_1(s_1, \widehat{\pi}_1(s_1)) \right) \\ &\quad - \left(\theta^\top \mathbf{r}_1(s_1, \widehat{\pi}_1(s_1)) + [\mathbb{P}_1 V_2^{\widehat{\pi}}](s_1, \widehat{\pi}_1(s_1); \theta) \right) \\ &= 2\widehat{u}_1(s_1, \widehat{\pi}_1(s_1)) + \left([\mathbb{P}_1 \widehat{V}_2](s_1, \widehat{\pi}_1(s_1)) - [\mathbb{P}_1 V_2^{\widehat{\pi}}](s_1, \widehat{\pi}_1(s_1); \theta) \right) \\ &= 2\widehat{u}_1(s_1, \widehat{\pi}_1(s_1)) + \mathbb{E}_{s_2 \sim \widehat{\pi}}[\widehat{V}_2(s_2) - V_2^{\widehat{\pi}}(s_2; \theta)] \\ &= \dots \\ &= 2\mathbb{E}_{\widehat{\pi}}\left[\sum_{h=1}^H \widehat{u}_h(s_h, a_h)\right] \\ &= 2V_1^{\widehat{\pi}}(s_1 | \widehat{u}),\end{aligned} \tag{24}$$

where (i) is uses Lemma 28. Therefore we have

$$\begin{aligned}V_1^*(s_1; \theta) - V_1^{\widehat{\pi}}(s_1; \theta) &\stackrel{(i)}{\leq} \widehat{V}_1(s_1) - V_1^{\widehat{\pi}}(s_1; \theta) \\ &\stackrel{(ii)}{\leq} 2V_1^{\widehat{\pi}}(s_1 | \widehat{u}) \\ &\stackrel{(iii)}{\leq} 2V_1^*(s_1 | \widehat{u}) \\ &= 2H \cdot V_1^*(s_1 | \widehat{u}/H) \\ &\stackrel{(iv)}{\leq} \mathcal{O}\left(\sqrt{d_{\text{lin}}^3 H^6 \epsilon^2 / K}\right) \\ &\stackrel{(v)}{\leq} \epsilon,\end{aligned}$$

where (i) uses Lemma 28, (ii) uses Equation 24, (iii) uses definition of optimal value function, (iv) uses Lemma 27, and (v) is due to $K \geq c_K[d_{\text{lin}}^3 H^6(\iota')^2/\epsilon^2]$ with a sufficiently large constant c_K ; It completes the proof. \square

E PROOF FOR SECTION 6

In this section we provide proofs and missing details for Section 6.

E.1 PROOF OF THEOREM 12

Define $\mathbf{v}^t = V_1^{\mu^t, \nu^t}(s_1)$ and note that $\mathbb{E}[\hat{\mathbf{v}}^t] = \mathbf{v}^t$.

Lemma 29. Define even E_3 to be:

$$\begin{cases} \|\frac{1}{T} \sum_{t=1}^T \mathbf{v}^t - \hat{\mathbf{v}}^t\| \leq \mathcal{O}(\sqrt{dH^2\iota/T}), \\ V_1^{\mu^t, \nu^t}(s_1; \boldsymbol{\theta}^t) - V_1^*(s_1; \boldsymbol{\theta}^t) \leq \epsilon/2 \quad \forall t \in [T]. \end{cases}$$

where $\iota = \log(d/\delta)$. We have $\mathbb{P}(E_0) \geq 1 - \delta$.

Proof of Lemma 29. We prove each claim holds with probability at least $1 - \delta/2$; applying union bound completes the proof.

First claim. Let \mathcal{F}_t be the filtration capturing all the randomness in the algorithm before iteration t . We have $\mathbb{E}[\hat{\mathbf{v}}^t | \mathcal{F}_t] = \mathbf{v}^t$ and we also know that $\|\hat{\mathbf{v}}^t\| \leq H$ almost surely. By applying Lemma 36, with probability at least $1 - \delta$ we have

$$\|\frac{1}{T} \sum_{t=1}^T \mathbf{v}^t - \hat{\mathbf{v}}^t\| \leq \mathcal{O}(\sqrt{H^2 \log[d/\delta]/T}),$$

which completes the proof.

Second claim. We have $K \geq m_{\text{RFE}}(\epsilon/2, \delta/2)$, therefore by probability at least $1 - \delta/2$ (Definition 10) we have

$$V_1^{\mu^t, \dagger}(s_1; \boldsymbol{\theta}^t) - V_1^{\dagger, \omega^t}(s_1; \boldsymbol{\theta}^t) \leq \epsilon/2,$$

Since (μ^t, ω^t) is the output of the planning phase. By definition of V^* , $V^{\cdot, \dagger}$, and $V^{\dagger, \cdot}$, we further know that

$$\begin{aligned} V_1^*(s_1; \boldsymbol{\theta}^t) &= \max_{\nu} V_1^{\dagger, \nu}(s_1; \boldsymbol{\theta}^t) \geq V_1^{\dagger, \omega^t}(s_1; \boldsymbol{\theta}^t) \\ V_1^{\mu^t, \dagger}(s_1; \boldsymbol{\theta}^t) &= \max_{\nu} V_1^{\mu^t, \nu}(s_1; \boldsymbol{\theta}^t) \geq V_1^{\mu^t, \nu^t}(s_1; \boldsymbol{\theta}^t) \end{aligned}$$

Combining the three equations gives us,

$$V_1^{\mu^t, \nu^t}(s_1; \boldsymbol{\theta}^t) - V_1^*(s_1; \boldsymbol{\theta}^t) \leq V_1^{\mu^t, \dagger}(s_1; \boldsymbol{\theta}^t) - V_1^{\dagger, \omega^t}(s_1; \boldsymbol{\theta}^t) \leq \epsilon/2,$$

and completes the proof. \square

Lemma 30. For any $\boldsymbol{\theta} \in \mathcal{B}(1)$, we have

$$V_1^*(s_1; \boldsymbol{\theta}) \leq \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle + \max_{\nu} \min_{\mu} \text{dist}(\mathbf{V}_1^{\mu, \nu}(s_1), \mathcal{C})$$

Proof of Lemma 30. Let $\alpha = \max_{\nu} \min_{\mu} \text{dist}(\mathbf{V}_1^{\mu, \nu}(s_1), \mathcal{C})$; therefore, for every max-player policy ν there exist a min-player policy $\bar{\mu}(\nu)$ such that $\text{dist}(\mathbf{V}_1^{\bar{\mu}(\nu), \nu}(s_1), \mathcal{C}) \leq \alpha$. Let $\Gamma_{\mathcal{C}}$ be the (Euclidean)

projection operator into \mathcal{C} . We have

$$\begin{aligned}
V_1^*(s_1; \boldsymbol{\theta}) &= V_1^{\mu^*, \nu^*}(s_1; \boldsymbol{\theta}) \\
&\leq V_1^{\bar{\mu}(\nu^*), \nu^*}(s_1; \boldsymbol{\theta}) \\
&= \langle \boldsymbol{\theta}, \mathbf{V}_1^{\bar{\mu}(\nu^*), \nu^*}(s_1) \rangle \\
&= \langle \boldsymbol{\theta}, \mathbf{V}_1^{\bar{\mu}(\nu^*), \nu^*}(s_1) - \Gamma_{\mathcal{C}}[\mathbf{V}_1^{\bar{\mu}(\nu^*), \nu^*}(s_1)] \rangle + \langle \boldsymbol{\theta}, \Gamma_{\mathcal{C}}[\mathbf{V}_1^{\bar{\mu}(\nu^*), \nu^*}(s_1)] \rangle \\
&\leq \|\boldsymbol{\theta}\| \text{dist}(\mathbf{V}_1^{\bar{\mu}(\nu^*), \nu^*}(s_1), \mathcal{C}) + \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle \\
&\leq \alpha + \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle,
\end{aligned}$$

Recalling that $\alpha = \max_{\nu} \min_{\mu} \text{dist}(\mathbf{V}_1^{\mu, \nu}(s_1), \mathcal{C})$ completes the proof. \square

Proof of Theorem 12. With probability at least $1 - \delta$, event E_3 (as in Definition 29) holds and we have

$$\begin{aligned}
\text{dist}\left(\frac{1}{T} \sum_{t=1}^T \mathbf{V}_1^{\mu^t, \nu^t}(s_1), \mathcal{C}\right) &= \text{dist}\left(\frac{1}{T} \sum_{t=1}^T \mathbf{v}^t, \mathcal{C}\right) \\
&\stackrel{(i)}{=} \max_{\boldsymbol{\theta} \in \mathcal{B}(1)} \left[\langle \boldsymbol{\theta}, \frac{1}{T} \sum_{t=1}^T \mathbf{v}^t \rangle - \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle \right] \\
&= \max_{\boldsymbol{\theta} \in \mathcal{B}(1)} \left[\frac{1}{T} \sum_{t=1}^T (\langle \boldsymbol{\theta}, \hat{\mathbf{v}}^t \rangle - \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle) + \langle \boldsymbol{\theta}, \frac{1}{T} \sum_{t=1}^T \mathbf{v}^t - \hat{\mathbf{v}}^t \rangle \right] \\
&\stackrel{(ii)}{\leq} \max_{\boldsymbol{\theta} \in \mathcal{B}(1)} \left[\frac{1}{T} \sum_{t=1}^T (\langle \boldsymbol{\theta}, \hat{\mathbf{v}}^t \rangle - \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}, \mathbf{x} \rangle) \right] + \mathcal{O}(\sqrt{dH^2\iota/T}) \\
&\stackrel{(iii)}{\leq} \frac{1}{T} \sum_{t=1}^T (\langle \boldsymbol{\theta}^t, \hat{\mathbf{v}}^t \rangle - \max_{\mathbf{x} \in \mathcal{C}} \langle \boldsymbol{\theta}^t, \mathbf{x} \rangle) + \mathcal{O}(\sqrt{H^2/T}) + \mathcal{O}(\sqrt{dH^2\iota/T}) \\
&\stackrel{(iv)}{\leq} \max_{\nu} \min_{\mu} \text{dist}(\mathbf{V}_1^{\mu, \nu}(s_1), \mathcal{C}) + \frac{1}{T} \sum_{t=1}^T (\langle \boldsymbol{\theta}^t, \hat{\mathbf{v}}^t \rangle - \mathbf{V}_1^*(s_1; \boldsymbol{\theta}^t)) + \mathcal{O}(\sqrt{dH^2\iota/T}) \\
&\stackrel{(v)}{\leq} \max_{\nu} \min_{\mu} \text{dist}(\mathbf{V}_1^{\mu, \nu}(s_1), \mathcal{C}) + \epsilon/2 + \frac{1}{T} \sum_{t=1}^T (\langle \boldsymbol{\theta}^t, \hat{\mathbf{v}}^t \rangle - \mathbf{V}_1^{\mu^t, \nu^t}(s_1; \boldsymbol{\theta}^t)) + \mathcal{O}(\sqrt{dH^2\iota/T}) \\
&= \max_{\nu} \min_{\mu} \text{dist}(\mathbf{V}_1^{\mu, \nu}(s_1), \mathcal{C}) + \epsilon/2 + \frac{1}{T} \sum_{t=1}^T \langle \boldsymbol{\theta}^t, \hat{\mathbf{v}}^t - \mathbf{v}^t \rangle + \mathcal{O}(\sqrt{dH^2\iota/T}) \\
&\stackrel{(vi)}{\leq} \max_{\nu} \min_{\mu} \text{dist}(\mathbf{V}_1^{\mu, \nu}(s_1), \mathcal{C}) + \epsilon/2 + \mathcal{O}(\sqrt{dH^2\iota/T}) \\
&\stackrel{(vii)}{\leq} \max_{\nu} \min_{\mu} \text{dist}(\mathbf{V}_1^{\mu, \nu}(s_1), \mathcal{C}) + \epsilon
\end{aligned}$$

where (i) is by Equation 6, (ii) is by first inequality in event E_3 together with Cauchy-Schwarz, (iii) is by guarantee of OGA in Theorem 15, (iv) is by Lemma 30, (v) is by second inequality in event E_3 , (vi) is by first inequality in event E_3 together with Cauchy-Schwarz, and finally (vii) is by setting $T \geq c(dH^2\iota/\epsilon^2)$ for large enough constant c , completing the proof. \square

E.2 PROOF OF THEOREM 13

E.2.1 ALGORITHM

Exploration phase. Similar to Algorithm 5, we use VI-Zero proposed by Liu et al. (2020) with different choice of hyperparameters. The pseudo-code is provided in Algorithm 8.

Planning phase. In the planning phase, given $\theta \in \mathcal{B}(1)$ as input we can use any planning algorithm for $\hat{\mathcal{G}}_\theta = (\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \hat{\mathbb{P}}^{\text{out}}, (\theta, \hat{\mathbf{r}}))$ where $\hat{\mathbf{r}}$ is empirical estimate of \mathbf{r} using collected samples $\{\mathbf{r}_h^k\}$. One such algorithm could be Nash value iteration (e.g. see Algorithm 5 in Liu et al. 2020) that computes Nash equilibrium policy for a *known* model.

Algorithm 8 VI-Zero for VMGs: Exploration Phase

```

1: Hyperparameters: Bonus  $\beta_t$ .
2: Initialize: for all  $(s, a, b, h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$ :  $\tilde{Q}_h(s, a, b) \leftarrow H$  and  $N_h(s, a, b) \leftarrow 0$ ,
3:   for all  $(s, a, b, h, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H] \times \mathcal{S}$ :  $N_h(s, a, b, s') \leftarrow 0$ ,
4:    $\Delta \leftarrow 0$ .
5: for episode  $k = 1, 2, \dots, K$  do
6:   for step  $h = H, H-1, \dots, 1$  do
7:     for state-action pair  $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$  do
8:        $t \leftarrow N_h(s, a, b)$ .
9:       if  $t > 0$  then
10:         $\tilde{Q}_h(s, a, b) \leftarrow \min\{\hat{\mathbb{P}}_h \tilde{V}_{h+1}\}(s, a, b) + \beta_t, H\}$ .
11:     for state  $s \in \mathcal{S}$  do
12:        $\tilde{V}_h(s) \leftarrow \max_{(a,b) \in \mathcal{A} \times \mathcal{B}} \tilde{Q}_h(s, a, b)$  and  $\pi_h(s) \leftarrow \operatorname{argmax}_{(a,b) \in \mathcal{A} \times \mathcal{B}} \tilde{Q}_h(s, a, b)$ 
13:     if  $\tilde{V}(s_1) \leq \Delta$  then
14:        $\Delta \leftarrow \tilde{V}(s_1)$  and  $\hat{\mathbb{P}}^{\text{out}} \leftarrow \hat{\mathbb{P}}_h$ 
15:     for step  $h = 1, 2, \dots, H$  do
16:       Take action  $(a_h, b_h) \leftarrow \pi_h(s_h)$  and observe next state  $s_{h+1}$ 
17:       Update  $N_h(s_h, a_h, b_h) \leftarrow N_h(s_h, a_h, b_h) + 1$ 
18:       Update  $N_h(s_h, a_h, b_h, s_{h+1}) \leftarrow N_h(s_h, a_h, b_h, s_{h+1}) + 1$ 
19:        $\hat{\mathbb{P}}_h(\cdot | s_h, a_h, b_h) \leftarrow N_h(s_h, a_h, b_h, \cdot) / N_h(s_h, a_h, b_h)$ 
20: Return  $\hat{\mathbb{P}}^{\text{out}}$ 

```

E.2.2 PROOF OF THEOREM 13

Proof is almost identical to proof of Theorem ?? provided in Appendix C; therefore, we only provide the statement for the main lemmas without proof.

Let $\hat{\mathbb{P}}^k$ and $\hat{\mathbf{r}}^k$ be our empirical estimates of the transition and the return vectors at the beginning of the k^{th} episode in Algorithm 8 and define $\hat{\mathcal{G}}^k = (\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \hat{\mathbb{P}}^k, \hat{\mathbf{r}}^k)$. We use $N_h^k(s, a, b)$ to denote the number of times we have visited state-action (s, a, b) in step h before k^{th} episode in Algorithm 8. We use superscript k to denote variable corresponding to episode k ; in particular, $(s_1^k, a_1^k, b_1^k, \dots, s_H^k, a_H^k, b_H^k)$ is the trajectory we have visited in the k^{th} episode.

For any $\theta \in \mathcal{B}(1)$, let $\hat{\mathcal{G}}_\theta^k$ be the scalarized VMG using vector θ (defined in Section 6). We use $\hat{V}^k(\cdot; \theta)$, $\hat{Q}^k(\cdot, \cdot, \cdot; \theta)$, and $(\hat{\mu}_\theta^k, \hat{\nu}_\theta^k) = (\hat{\mu}^k(\cdot; \theta), \hat{\nu}^k(\cdot; \theta))$ to denote the optimal value function, optimal Q-value function, and Nash equilibrium policy of $\hat{\mathcal{G}}_\theta^k$ respectively. Therefore, we have

$$\begin{aligned}
\hat{Q}_h^k(s, a, b; \theta) &= [\hat{\mathbb{P}}_h^k \hat{V}_{h+1}^k](s, a, b; \theta) + \hat{r}_h^k(s, a, b; \theta), \\
\hat{V}_h^k(s; \theta) &= \min_{\mu} \max_{\nu} [\mathbb{D}_{\mu \times \nu} \hat{Q}_h^k](s; \theta), \\
\hat{V}_h^k(s; \theta) &= [\mathbb{D}_{\hat{\mu}_\theta^k \times \hat{\nu}_\theta^k} \hat{Q}_h^k](s; \theta).
\end{aligned} \tag{25}$$

Theorem 31 (restatement of Theorem 13). *There exist absolute constants c_β and c_K , such that for any $\epsilon \in (0, H]$, $\delta \in (0, 1]$, if we choose bonus $\beta_t = c_\beta(\sqrt{\min\{d, S\}}H^2\iota/t + H^2S\iota/t)$ where $\iota = \log[dSABKH/\delta]$, and run the exploration phase (Algorithm 8) for $K \geq c_K(\min\{d, S\}H^4SAB\iota'/\epsilon^2 + H^3S^2AB(\iota')^2/\epsilon)$ episodes where $\iota' = \log[dSABH/(\epsilon\delta)]$, then with probability at least $1 - \delta$, the algorithm satisfies for all $\theta \in \mathcal{B}(1)$*

$$V_1^{\mu_\theta, \dagger}(s_1; \theta) - V_1^{\dagger, \nu_\theta}(s_1; \theta) = [V_1^{\mu_\theta, \dagger}(s_1; \theta) - V_1^*(s_1; \theta)] + [V_1^*(s_1; \theta) - V_1^{\dagger, \nu_\theta}(s_1; \theta)] \leq \epsilon,$$

where (μ_θ, ν_θ) is the output of any planning algorithm (e.g., Nash value iteration) for the Markov game $\widehat{\mathcal{G}}_\theta^{\text{out}}$. Therefore, we have

$$m_{\text{RFE}}(\epsilon, \delta) \leq \mathcal{O}\left(\frac{\min\{d, S\}H^4SAB\iota'}{\epsilon^2} + \frac{H^3S^2AB(\iota')^2}{\epsilon}\right).$$

The bonus for episode k can be written as

$$\beta_h^k(s, a, b) = c_\beta \left(\sqrt{\frac{\min\{d, S\}H^2\iota}{\max\{N_h^k(s, a, b), 1\}}} + \frac{H^2S\iota}{\max\{N_h^k(s, a, b), 1\}} \right), \quad (26)$$

where $\iota = \log[dSABKH/\delta]$ and c_β is some large absolute constant.

We start with the concentration lemma similar to Lemma 20.

Lemma 32. *Let c be some large absolute constant. Define event E_4 to be: for all $(s, a, b, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times \mathcal{S} \times [H]$, $k \in [K]$, and $\theta \in \mathcal{B}(1)$,*

$$\begin{cases} |[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^*](s, a, b; \theta)| & \leq c\sqrt{\frac{\min\{d, S\}H^2\iota}{\max\{N_h^k(s, a, b), 1\}}}, \\ |(\widehat{r}_h^k - r_h)(s, a, b; \theta)| & \leq c\sqrt{\frac{\iota}{\max\{N_h^k(s, a, b), 1\}}}, \\ |(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(s' | s, a, b)| & \leq c\left(\sqrt{\frac{\widehat{\mathbb{P}}_h^k(s' | s, a, b)\iota}{\max\{N_h^k(s, a, b), 1\}}} + \frac{\iota}{\max\{N_h^k(s, a, b), 1\}}\right), \end{cases} \quad (27)$$

where $\iota = \log[dSABKH/\delta]$. We have $\mathbb{P}(E_4) \geq 1 - \delta$.

Similar to Lemma 21, the following lemma shows that the optimal value functions of $\widehat{\mathcal{G}}_\theta^k$ are close to the optimal value functions of \mathcal{G}_θ and their difference is controlled by \widetilde{Q} and \widetilde{V} computed in Algorithm 8.

Lemma 33. *Suppose event E_4 holds (defined in Lemma 32); then, for all $(s, a, b, k, h, \theta) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [K] \times [H] \times \mathcal{B}(1)$ we have*

$$\begin{aligned} |\widehat{Q}_h^k(s, a, b; \theta) - Q_h^*(s, a, b; \theta)| & \leq \widetilde{Q}_h^k(s, a, b), \\ |\widehat{V}_h^k(s; \theta) - V_h^*(s; \theta)| & \leq \widetilde{V}_h^k(s). \end{aligned} \quad (28)$$

Similar to Lemma 22, now we are ready to introduce the main lemma that shows value of $\widehat{\pi}_\theta^k$ under the true model is close to its value under empirical model. The difference is controlled by \widetilde{Q} and \widetilde{V} computed in Algorithm 8.

Lemma 34. *Suppose event E_4 holds (defined in Lemma 32); then, for all $(s, a, b, k, h, \theta) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [K] \times [H] \times \mathcal{B}(1)$ we have*

$$\begin{aligned} |\widehat{Q}_h^k(s, a, b; \theta) - Q_h^{\dagger, \widehat{V}_h^k}(s, a, b; \theta)| & \leq \alpha_h \widetilde{Q}_h^k(s, a, b), \\ |\widehat{V}_h^k(s; \theta) - V_h^{\dagger, \widehat{Q}_h^k}(s; \theta)| & \leq \alpha_h \widetilde{V}_h^k(s), \end{aligned} \quad (29)$$

and

$$\begin{aligned} |\widehat{Q}_h^k(s, a, b; \theta) - Q_h^{\widehat{\mu}_\theta^k, \dagger}(s, a, b; \theta)| & \leq \alpha_h \widetilde{Q}_h^k(s, a, b), \\ |\widehat{V}_h^k(s; \theta) - V_h^{\widehat{\mu}_\theta^k, \dagger}(s; \theta)| & \leq \alpha_h \widetilde{V}_h^k(s), \end{aligned} \quad (30)$$

where $\alpha_{H+1} = 1$ and $\alpha_h = [(1 + \frac{1}{H})\alpha_{h+1} + \frac{1}{H}]$; we have $1 \leq \alpha_h \leq 5$ for $h \in [H]$.

Similar to Lemma 23, we can bound the uncertainty using the following lemma.

Theorem 35. *For any $\delta \in (0, 1]$, if we choose β_t^k in Algorithm 8 as in Equation 26; then, with probability at least $1 - \delta$, we have*

$$\sum_{k=1}^K \widetilde{V}_1^k(s_1) \leq \mathcal{O}(\sqrt{\min\{d, S\}H^4SABK\iota} + H^3S^2AB\iota^2).$$

Proof of Theorem 31 (restatement of Theorem 13). By Algorithm 8, we have $\text{out} = \arg\min_{k \in [K]} \tilde{V}_1^k(s_1)$, resulting in $\hat{V}_1^{\text{out}}(s_1) \leq \frac{1}{K} \sum_{k=1}^K \tilde{V}_1^k(s_1)$. Therefore, with probability at least $1 - 2\delta$, for any vector $\theta \in \mathcal{B}(1)$ we have

$$\begin{aligned}
V_1^{\hat{\mu}_{\theta}^{\text{out}, \dagger}}(s_1; \theta) - V_1^{\dagger, \hat{\nu}_{\theta}^{\text{out}}}(s_1; \theta) &\leq |V_1^{\hat{\mu}_{\theta}^{\text{out}, \dagger}}(s_1; \theta) - \hat{V}_1^{\text{out}}(s_1; \theta)| + |\hat{V}_1^{\text{out}}(s_1; \theta) - V_1^{\dagger, \hat{\nu}_{\theta}^{\text{out}}}(s_1; \theta)| \\
&\stackrel{(i)}{\leq} 2\alpha_1 \tilde{V}_1^{\text{out}}(s_1) \\
&\leq 10\tilde{V}_1^{\text{out}}(s_1) \\
&\leq \frac{10}{K} \sum_{k=1}^K \tilde{V}_1^k(s_1) \\
&\stackrel{(ii)}{\leq} \mathcal{O}(\sqrt{\min\{d, S\}H^4SAB\iota/K} + H^3S^2AB\iota^2/K) \\
&\stackrel{(iii)}{\leq} \epsilon,
\end{aligned}$$

where (i) is due to Lemma 34, (ii) is due to Theorem 35, and (iii) is due to $K \geq c_K(\min\{d, S\}H^4SAB\iota'/\epsilon^2 + H^3S^2AB(\iota')^2/\epsilon)$ with a sufficiently large constant c_K . Rescaling δ completes the proof. \square

F AUXILIARY TOOLS

Lemma 36 (Hoeffding type inequality for norm-subGaussian, Corollary 7 in Jin et al. 2019). *Let $\{\mathbf{X}_t\}_{t \in [T]}$ be a d -dimensional vector-valued random variable. Consider filtration $\{\mathcal{F}_t\}_{t \in [T]}$ and define $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_t]$. If $\|\mathbf{X}_t\| \leq R$ almost surely, then it holds with probability at least $1 - \delta$,*

$$\left\| \sum_{t=1}^T \mathbf{X}_t - \mathbb{E}_{t-1}[\mathbf{X}_t] \right\| \leq \mathcal{O}(R\sqrt{T \log[d/\delta]}).$$